Scene Understanding for Intelligent Transportation and Mobility Assistance Systems

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Jiaming Zhang
aus Guangdong, China

Tag der mündlichen Prüfung:	13. November 2023
Hauptreferent:	Prof. Dr.-Ing. Rainer Stiefelhagen Karlsruher Institut für Technologie
Korreferent:	Prof. Philip H.S. Torr University of Oxford

Jiaming Zhang: *Scene Understanding for Intelligent Transportation and Mobility Assistance Systems*

ABSTRACT

Scene understanding – a process of parsing visual perception of surrounding environment into human-readable information – can be processed via pixel-wise image semantic segmentation, supporting intelligent systems to make correct decisions when interacting with the environment. This versatile approach finds extensive applications, such as autonomous vehicles and assistive technology. *As estimated in 2020, there are approximately 596 million people worldwide suffering from distance vision impairment, of whom 43 million were blind.* Thus, it is meaningful to extend the benefit of scene understanding techniques to include People with Visual Impairments (PVI). In this thesis, we mainly investigate scene understanding in two distinct yet correlated research fields: *Intelligent Transportation Systems (ITS)* and *Mobility Assistance Systems (MAS).*

In the domain of ITS, our research focuses on two aspects. (1) *Panoramic* semantic segmentation, which entails pixel-level parsing of 360° driving scenes, yields a *omnidirectional* scene understanding. For the first time, we establish a new dataset (*DensePASS*) with 360° semantic annotations for benchmarking pinhole-to-panoramic domain adaptation. Besides, our unsupervised *Trans4PASS* model maintains comparable performance to fully-supervised state-of-the-arts. (2) *Multimodal* semantic segmentation is able to improve *robustness* by incorporating complementary modalities. As another contribution, we create a new benchmark (*DeLiVER*) for enabling arbitrary-modal semantic segmentation. Moreover, we introduce unified models *CMX* and *CMNeXt* for conducting RGB-X semantic segmentation, yielding a flexible and seamless fusion of RGB with Depth, Event, LiDAR, Thermal, Polarization, and Light-field images.

In the field of MAS, our research focus lies in developing scene understanding and assistance systems for both drivers and pedestrians, while also considering PVI. Our contributions are two-fold. (3) Vision localization and semantic mapping are thoroughly explored to empower *navigational* scene understanding. A novel feature matching model *MatchFormer* is implemented to perform robust pose estimation indoors and outdoors. To enhance map accessibility for pre-exploring destinations, two frameworks *Trans4Map* and *360BEV* are proposed for indoor semantic mapping using sequences or single images. (4) To develop practical applications of MAS, we explore *generalizable* scene understanding for transparent object segmentation and adverse scene segmentation. Glass-like objects present architectural obstacles that impede the mobility of PVI. To handle the *safety-critical* cases, we leverage our *Trans4Trans* model to design a wearable *Vision4Blind* system and iteratively improve it based on user feedback. Furthermore, a proof-of-concept prototype, *i.e.*, “flying guide dog”, is implemented to assess new possibilities in scene understanding and assistance systems.

By exploring two interconnected research fields, this thesis has unfolded significant insights into scene understanding – including new datasets created to facilitate the community, advanced models to enhance perception, and wearable systems designed to assist PVI – contributing to the advancement and realization of intelligent mobility.

ZUSAMMENFASSUNG

Szenenverständnis – ein Prozess der Analyse der visuellen Wahrnehmung der Umgebung in menschenlesbare Informationen – kann durch pixelweise Bildsemantiksegmentierung verarbeitet werden und intelligente Systeme unterstützen, um bei der Interaktion mit der Umgebung richtige Entscheidungen zu treffen. Diese vielseitige Methode findet vielfältige Anwendungen, wie z. B. autonome Fahrzeuge und Assistenztechnologie. *Nach Schätzungen aus dem Jahr 2020 leiden weltweit rund 596 Millionen Menschen an Sehbehinderungen, von denen 43 Millionen blind waren.* Daher ist es sinnvoll, die Vorteile des Szenenverständnisses auch für Menschen mit Sehbehinderung zu erweitern. In dieser Dissertation untersuchen wir das Szenenverständnis in zwei unterschiedlichen, aber miteinander verbundenen Forschungsbereichen: *Intelligente Transportsysteme* und *Mobilitätsassistenzsysteme.*

Im Kontext der intelligenten Transportsysteme konzentriert sich unsere Forschung auf zwei Schlüsselaspekte. (1) *Panoramische* semantische Segmentierung, das die pixelgenaue Analyse von 360°-Fahrszenen beinhaltet, ergibt ein *holistisches* Szenenverständnis. Erstmals erstellen wir ein neues Dataset (*DensePASS*) mit 360°-semantischen Annotationen für die Benchmarking von pinhole-to-panoramic Domain Adaptation. Darüber hinaus verfügt unser unüberwachtes *Trans4PASS*-Modell über eine vergleichbare Leistung zu vollüberwachten modernsten Modellen. (2) *Multimodale* semantische Segmentierung kann die *Robustheit* verbessern, indem komplementäre Modalitäten integriert werden. Als weiteren Beitrag erstellen wir einen neuen Benchmark (*DeLiVER*) für beliebig modale semantische Segmentierung. Darüber hinaus stellen wir die einheitlichen Modelle *CMX* und *CMNeXt* für RGB-X-semantische Segmentierung vor, die eine flexible und nahtlose Fusion von RGB mit Depth, Event, LiDAR, Thermal, Polarization und Light-field-Bildern ermöglichen.

Im Bereich Mobilitätsassistenzsysteme liegt unser Forschungsschwerpunkt auf der Entwicklung von Szenenverständnis- und Assistenzsystemen für Fahrer, Fußgänger und Menschen mit Sehbehinderung. Unsere Beiträge sind zweifach. (3) Visionslokalisierung und semantische Kartierung werden umfassend erforscht, um das *navigationale* Szenenverständnis zu unterstützen. Ein neuartiges Feature-Matching-Modell *MatchFormer* wird implementiert, um robuste Pose-Estimation in Innen- und Außenbereichen durchzuführen. Um die Map-Barrierefreiheit für die Vorerkundung des Zielorts zu verbessern, werden zwei Frameworks *Trans4Map* und *360BEV* für indoor semantische Kartierung mit Sequenzen oder Einzelbildern vorgeschlagen. (4) Um praktische Anwendungen von Mobilitätsassistenzsystemen zu entwickeln, erforschen wir das *generalisierbare* Szenenverständnis durch transparente Objektsegmentierung und adverse Szenensegmentierung. Glasähnliche Objekte stellen architektonische Hindernisse dar, die die Mobilität von Menschen mit Sehbehinderung behindern. Um *sicherheitskritische* Fälle zu bewältigen, nutzen wir unser *Trans4Trans*-Modell, um ein tragbares *Vision4Blind*-System zu entwerfen und es durch Benutzerfeedback iterativ zu verbessern. Darüber hinaus wird ein Proof-of-Concept-Prototyp, d. h. ein „flying

guide dog“, implementiert, um neue Möglichkeiten im Bereich Szenenverständnis und Assistenzsysteme zu bewerten.

Durch die Erforschung von zwei miteinander verbundenen Forschungsbereichen hat diese Dissertation wichtige Erkenntnisse zum Szenenverständnis gewonnen – einschließlich neuer Datensätze zur Erleichterung der Community, fortschrittlicher Modelle zur Verbesserung der Wahrnehmung und tragbarer Systeme zur Unterstützung von Menschen mit Sehbehinderungen – und so zur Weiterentwicklung und Verwirklichung intelligenter Mobilität beigetragen.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to all those who have contributed to my dissertation, including my advisors, collaborators, colleagues, friends and family. Without their support and encouragement, this journey would not be possible.

First, I extend my deepest appreciation to my supervisor, Prof. Rainer Stiefelhagen, for his unwavering support, valuable insights, and encouragement throughout this research. His expertise and dedication have been instrumental in shaping the direction of this research. In addition to providing professional insights, he has also granted me the freedom to explore new ideas and topics, which serves as a great source of encouragement and belief in my abilities.

I am immensely grateful to my second advisor, Prof. Philip Torr, for hosting me during my visit to the University of Oxford and providing valuable suggestions on this thesis and my doctoral research. It is a great honor to work with him and his renowned Torr Vision Group, pioneers in the field of scene understanding.

I want to express my sincere appreciation for the members of my dissertation committee, Prof. Christian Wressnegger, Prof. Hannes Hartenstein, and Prof. Ina Schaefer, for their time and effort dedicated to this project. Besides, I want to express my sincere appreciation for all suggestions received during the promotion talks with Prof. Frank Bellosa, Prof. Dorothea Wagner, and Prof. Tamim Asfour.

My heartfelt thanks go out to all collaborators at CV:HCI lab, Kailun, Kunyu, Junwei, Ruiping, Yufan, Alina, Simon, Constantin, Alexander, David, Zdravko, Manuel, and Saquib, for their valuable discussions. And a big thanks to Corinna, who always kindly helps me with organization stuff. Also, I am deeply grateful to colleagues at ACCESS@KIT lab, Karin, Thorsten, Gerhard, Christina, Omar, Angela, and Julia, for their constructive feedback and suggestions.

Words cannot adequately express my gratitude to my family and friends for their unwavering encouragement and emotional support. I am profoundly thankful to my family, my mother Qilian, my wife Xiaoxi, my sisters Limei, Meiling and Yuting, and my soon-to-arrive baby. All of your support and belief in me kept me motivated during the challenging phases of this research. As a special thank you to my father, I am extremely grateful. Besides, also many thanks to my friends Jie, Boyang and more, for their friendship and relaxing chats.

I would also like to acknowledge the support provided by Helmholtz Association, KIT, BMAS and DAAD, for their financial assistance to carry out this research.

Last but not least, I would like to express my profound gratitude to all the blind and visually impaired participants who generously shared their time, knowledge, suggestions and comments. Your contributions made this study and research possible. Thank you all for being part of this project and for making it a truly rewarding experience.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND

1 INTRODUCTION

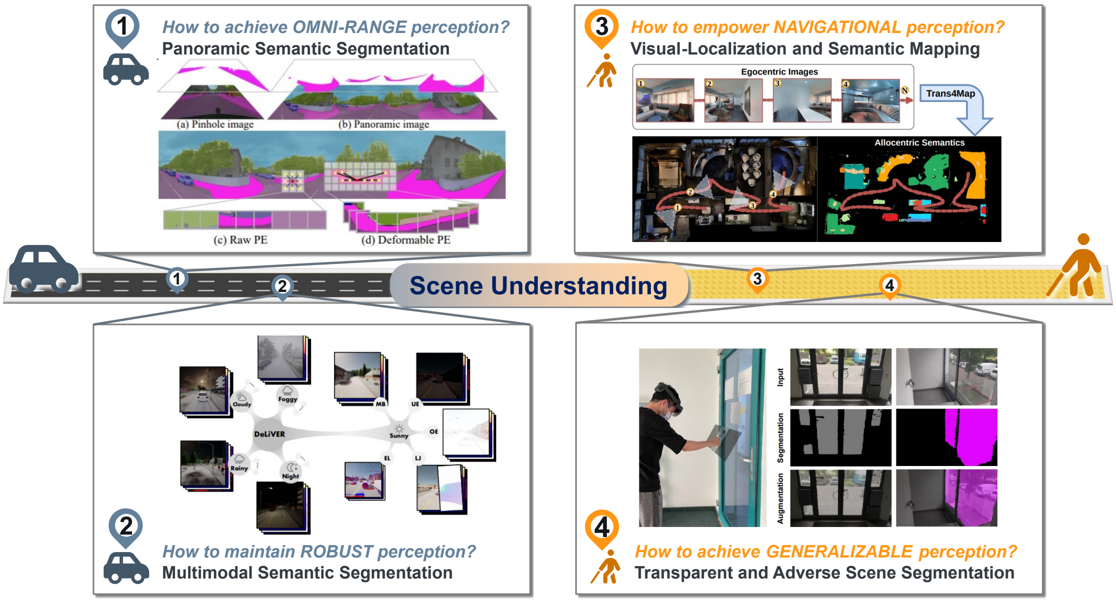
> Scene understanding is a process of interpreting visual data from the environment to identify objects, perceive spatial relationships, and comprehend contextual information. As illustrated in Figure 1, this dissertation is driven by a dual-pronged exploration of two distinct yet interrelated research domains, *i.e.*, *Intelligent Transportation Systems (ITS)* and *Mobility Assistance Systems (MAS)*. This synergistic development is dedicated to shaping an intelligent mobility system that benefits all traffic participants, from drivers and pedestrians to People with Visual Impairments (PVI).

1.1 SCENE UNDERSTANDING

Imagine a smart city in the future, where autonomous vehicles weave seamlessly through the urban environment, effortlessly recognizing pedestrians and traffic lights, and easily navigate complex intersections. All the while, pedestrians can move smoothly between destinations, and even people with disabilities can move safely and independently. Such a remarkable intelligent mobility utopia – that is more efficient, accessible and livable for everyone – is being progressively realized through the development of advanced technologies, showing the next-generation transportation – one that depends on the capabilities of scene understanding techniques.

Recently, scene understanding has emerged as a pivotal research area in the computer vision community, given its role in translating visual perceptions from the surrounding into meaningful and human-readable insights. This process is facilitated through pixel-wise image semantic segmentation, whereby every pixel in an image is analyzed and identified to provide intelligent systems with the crucial information. The versatility of scene understanding is impacting a wide range of downstream applications, such as autonomous vehicles and intelligent transportation systems. Moreover, extending the benefits of scene understanding to include People with Visual Impairments (PVI), who often encounter challenges in using transportation infrastructures, is meaningful and essential, because it addresses the impediments to their mobility independence. *As estimated in 2020, there are approximately 596 million people worldwide suffering from distance vision impairment, of whom 43 million were blind* [15]. Given their status as the most important and vulnerable road users, incorporating the perspectives of PVI is crucial in the development of intelligent mobility systems.

Therefore, in pursuit of the intelligent mobility utopia for all traffic participants, this thesis mainly delves into image semantic segmentation, aiming to provide advanced scene understanding approaches for both *Intelligent Transportation Systems (ITS)* and *Mobility Assistance Systems (MAS)*, as shown in Figure 1.


1
How to achieve OMNI-RANGE perception?
Panoramic Semantic Segmentation
(a) Pinhole image
(b) Panoramic image
(c) Raw PE
(d) Deformable PE
3
How to empower NAVIGATIONAL perception?
Visual-Localization and Semantic Mapping
Egocentric Images
1
2
3
4
N
Trans4Map
Allocentric Semantics
Scene Understanding
1
2
3
4
Cloudy
Foggy
MB
UE
DeLiVER
Sunny
OE
Rainy
Night
EL
LJ
2
How to maintain ROBUST perception?
Multimodal Semantic Segmentation
Input
Segmentation
Augmentation
4
How to achieve GENERALIZABLE perception?
Transparent and Adverse Scene Segmentation

(a) Intelligent Transportation Systems (b) Mobility Assistance Systems

Figure 1: The summary of the research questions and contributions in both domains.

1.1.1 *In the Domain of Intelligent Transportation Systems*

The evolution of transportation systems has witnessed a remarkable change towards intelligent and autonomous solutions. As depicted in Figure 1a, through scene understanding, *Intelligent Transportation Systems (ITS)* utilizes the power of advanced image segmentation techniques to decode the visual complexities of the traffic environment. This enables vehicles to perceive and respond intelligently to their environment, such as understanding the intricate dynamics of driving scenes, recognizing the state of traffic control devices, and predicting the category of other entities.

However, the driving scene in the real world is more complex and highly dynamic. One of the most significant challenges for ITS is the recognition and understanding of all objects in the driving scene. There are many different objects and entities that can be present within the traffic scene, including vehicles, pedestrians, cyclists, traffic signs, and lights [44]. These objects vary in terms of sizes, shapes, and colors, and they can be moving in different directions and interacting with one another. Another challenge for ITS is the dynamic behavior of road users. Pedestrians, cyclists, and even drivers can make sudden and unexpected movements. For instance, a pedestrian might dash across an intersection within the blind spot of a vehicle [285]. This complexity compounds the difficulty of accurately recognizing and understanding all objects.

Omnidirectional scene understanding of the entire driving scene is a key technique that can help to address the aforementioned challenges. To achieve this, panoramic semantic segmentation (in Figure 1a - ❶), powered by a single panoramic camera with 360-degree field of view (FoV), is an applicable approach. This streamlined setup in autonomous vehicles minimizes hardware complexity, installation, and calibration efforts, while still providing an omni-range view of the driving scenario. In essence, the combination of 360-degree cameras and image semantic segmentation yields a omni-range and seamless perception solution, enabling ITS to make more informed decisions about the safe and efficient operation of autonomous vehicles.

Apart from the impact of FoV, other various influencing factors may also be encountered during the practical driving. For example, the driving scene is frequently obscured by adverse weather conditions, such as fog, rain, or snow, which can significantly degrade the quality of images and sensor data, making it difficult for ITS systems to accurately perceive the road environment [171]. Besides, even in ideal weather conditions, sensor failures can also pose a challenge to ITS systems. For instance, if a camera or LiDAR sensor fails, the system will lose one of its main sources of information about the traffic surroundings. This can lead to a large drop in perception performance, as the system will have to rely on less data to make decisions.

Robust scene understanding stands as a vital point in addressing these challenges, which not only enhances ITS systems against the vulnerabilities posed by sensor failures but also augments their capacity to navigate through adverse traffic conditions. To attain this goal, multimodal semantic segmentation (in Figure 1a - ❷) emerges as a novel and promising solution, which can fuse distinct data modalities and excavate complementary information. For instance, *depth* measurements serve to help identify the boundary of objects and offer geometric information of dense scene elements [36, 82, 278]. Moreover, the integration and cooperation of multiple sensors holds the potential to effectively combat individual sensor failures [279].

Together, *panoramic* semantic segmentation and *multimodal* semantic segmentation can be jointly explored to improve the performance of ITS. Panoramic semantic segmentation enhances scene understanding through omnidirectional perception of the driving environment, while multimodal semantic segmentation improves robust scene understanding by fusing diverse sensory data sources.

1.1.2 *In the Field of Mobility Assistance Systems*

Beyond the transportation perspective, another vital factor of the intelligent mobility utopia is *Mobility Assistance Systems (MAS)*, which enhance pedestrian experiences and ensure accessible mobility for people with diverse needs, including those with visual impairments. As shown in Figure 1b, the combination of scene understanding and assistive technology becomes a potent support to achieve this feature. Solving mobility barriers encountered by pedestrians and the visually impaired contributes significantly to improve accessibility and intelligence.

Navigating independently is one of the challenges for individuals with visual impairments, underscoring the important role of MAS in improving their mobility. Be it walking around familiar or unfamiliar terrain, indoors or outdoors, the ability to move confidently between destinations remains a fundamental aspiration, since they may have difficulty in locating their positions in unfamiliar environments, finding their way around, avoiding obstacles ahead, and understanding their surroundings.

Navigational scene understanding is a crucial component to empower the navigational perception feature in MAS (in Figure 1b - ❸). On the one hand, visual localization allows MAS to determine its current position in the environment based on visual inputs [40]. This information is essential for the MAS to know the orientation and direction of the user. On the other hand, semantic mapping applied to MAS can understand the environment by assigning semantic labels to different objects in the given visual inputs. This information can be used to plan and execute navigation routes, as well as to provide users with holistic understanding about their surroundings in advance, so as to improve their accessibility.

In addition to navigation, many mobility obstacles must be addressed in the practical usage of assistance perception systems. One notable example is the recognition of glazed facades and transparent objects, which is seldom addressed by existing perception systems. Understanding the intricacies of glass architecture [16] and glass doors [138, 141] is particularly importance for both sighted and blind people. Transparent objects frequently present architectural obstacles that impede the mobility of those with low vision or blindness. For instance, imagine a scenario where an assistive system inaccurately perceives a path behind a closed glass door as accessible, potentially leading to missteps and endangering the user.

Generalizable scene understanding enables image segmentation models to handle adverse and corner cases when using in the real-world scenarios. For example, transparent object segmentation can be used to adaptively identify glass-like objects when walking indoors or outdoors (in Figure 1b - ④). With this feature, MAS can correctly identify safety-critical transparent objects, such as glass doors and windows. This information is important because transparent objects can be difficult to be aware of by people with visual impairments and sighted people. By accurately recognizing corner cases, MAS can help to keep users safe by avoiding collisions and ensuring that they understand their surroundings and how to interact with them correctly.

Therefore, by exploring both *navigational* scene understanding and *generalizable* scene understanding, mobility assistance systems (MAS) can be empowered to overcome the real-world challenges faced by pedestrians and those with visual impairments, contributing to a more inclusive and accessible mobility.

1.2 MOTIVATION AND GOALS

The main motivation of this thesis is to explore the *empowerment* of scene understanding in constructing an intelligent mobility utopia where autonomous vehicles glide effortlessly through intricate thoroughfares, pedestrians traverse with ease, and individuals with disabilities navigate safely and independently. Specially, the scene understanding methods will be mainly investigated in two fields of *intelligent transportation systems* and *mobility assistance systems*, which can assist both drivers and pedestrians, as well as people with visual impairments.

The goal of this thesis is to propose novel and effective solutions for scene understanding that can improve the perception ability of intelligent vehicles and assist people with visual impairments to navigate around safely and independently. Specifically, in the field of ITS, *omnidirectional* and *robust* perception ability can be achieved through promising *panoramic* and *multimodal* semantic segmentation. This will allow intelligent vehicles to better and stably understand their surroundings. In the domain of MAS, *navigational* and *generalizable* scene understanding are mainly explored to empower the independent navigation and safety-critical object recognition. Besides, human-friendly wearable systems and applications are implemented through image semantic segmentation and evaluated with the target group.

1.3 RESEARCH QUESTIONS

The main research objective of this thesis is to explore *scene understanding* to achieve intelligent mobility for all traffic participants, encompassing not only the perspective of drivers but also that of pedestrians, particularly those with visual impairments. As illustrated in Figure 1, scene understanding is investigated in two distinct yet interrelated domains, *i.e.*, *Intelligent Transportation Systems (ITS)* (Figure 1a) and *Mobility Assistance Systems (MAS)* (Figure 1b).

In the context of ITS, there are two key research questions.

RQ 1. ***How to achieve consistent perception in all directions at once?*** Omnidirectional perception can consistently understand the entire scene, including the layout of roads, the perception of other vehicles, pedestrians, cyclists, and traffic control devices. The challenges include that the environment around an intelligent vehicle is constantly changing and the interaction behaviors from other entities are less predictable. 360° cameras offer an omni-range view via a single sensor. Based on that, panoramic semantic segmentation is a potential solution to address the challenges. However, the scarcity of annotated panoramic images presents another difficulty in learning panoramic segmentation.

RQ 2. ***How to stabilize scene perception in a unified manner?*** Robust perception is to understand the environment even in the presence of noise and uncertainty. The challenges of system robustness encompass diverse influential factors, such as adverse weather conditions and sensor failure cases. To stabilize perception, fusing different sensors is a potential solution to provide more data to manage adverse scenarios, as well as to counter sensor failures. However, implementing a unified multimodal fusion method is crucial yet challenging to combine and extract the complementary information from various sensory modalities.

In the field of MAS, the two main research questions are listed.

RQ 3. ***How to advance on-site navigational perception and beforehand map accessibility?*** Visual localization and semantic mapping are two fundamental components that can be used to empower independent mobility for people with visual impairments. Visual localization is crucial for estimating location in on-site navigation, while semantic mapping can provide users with an accessible map to understand their destination in advance. However, due to the textureless indoor scenes, the indoor visual localization and semantic mapping is still a challenge in the field of navigational scene understanding.

RQ 4. ***How to handle corner cases and adverse situations in real-world applications?*** In practical scenarios, safety-critical cases are situations that pose a risk to the safety of users. For instance, glass doors and windows, which often lack distinct texture yet are commonplace both indoors and outdoors, can be particularly challenging to detect. To address these hazards, generalizable scene understanding is a way to enable general models to handle uncommon cases, such as transparent object segmentation to address glass-like scenes. By considering the synergy of walking and driving views, methods tailored for corner cases can be adapted to address adverse driving scenes as well.

1.4 THESIS OUTLINE

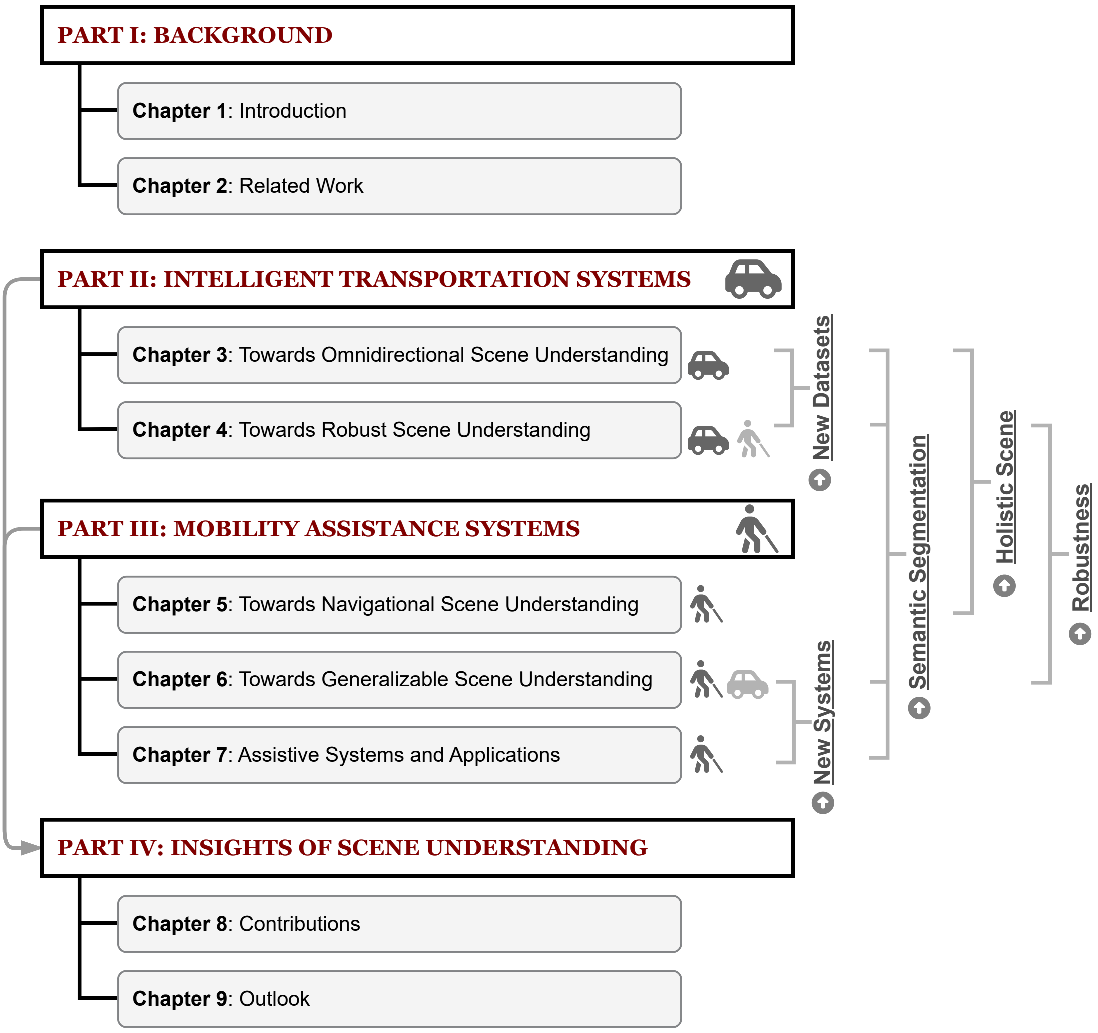

PART I: BACKGROUND
Chapter 1: Introduction
Chapter 2: Related Work
PART II: INTELLIGENT TRANSPORTATION SYSTEMS
Chapter 3: Towards Omnidirectional Scene Understanding
Chapter 4: Towards Robust Scene Understanding
PART III: MOBILITY ASSISTANCE SYSTEMS
Chapter 5: Towards Navigational Scene Understanding
Chapter 6: Towards Generalizable Scene Understanding
Chapter 7: Assistive Systems and Applications
PART IV: INSIGHTS OF SCENE UNDERSTANDING
Chapter 8: Contributions
Chapter 9: Outlook
New Datasets
Semantic Segmentation
Holistic Scene
Robustness
New Systems

Figure 2: The outline of the dissertation and the relationship between chapters.

The thesis outline is presented in Figure 2. There are 4 parts and 9 chapters. PART I presents the background of scene understanding with introduction (Chapter 1) and related work (Chapter 2). PART II focuses on panoramic semantic segmentation (Chapter 3) and multimodal semantic segmentation (Chapter 4). PART III involves navigational (Chapter 5) and generalizable (Chapter 6) scene understanding, and additional proof-of-concept prototypes (Chapter 7). PART IV includes contributions (Chapter 8) of this research and outlook (Chapter 9) about future work.

Besides, Figure 2 shows the relationship between chapters. While methods in PART II are proposed for *Intelligent Transportation Systems* (ITS marked as 🚗), approaches in PART III are mainly for *Mobility Assistance Systems* (MAS marked as 🧑‍🦯). We found that some methods are mutually applicable in both domains, which are marked as light-gray 🧑‍🦯 and 🚗. Chapter 3, 4, 6, and 7 are related to semantic segmentation. Chapter 3 and 5 are aiming at holistic perception. Chapter 4 and 6 are relevant to robustness. Two new datasets are presented in Chapter 3 and 4, while new systems (*e.g.*, the *Vision4Blind* system) are evaluated with user studies in Chapter 6 and 7.

PART I: BACKGROUND

Chapter 2: Related Work. This chapter provides a review of the existing literature and previous research works related to the field of scene understanding. The most relevant settings of image semantic segmentation are studies and presented, including the state-of-the-art methods and ideas. Besides, a detailed list of benchmarks for semantic segmentation is divided in to four different perspectives. Moreover, the recent and novel vision-based assistance systems are investigated and presented to provide preliminary knowledge and insights about assistive technology.

PART II: INTELLIGENT TRANSPORTATION SYSTEMS (ITS)

Chapter 3: Towards Omnidirectional Scene Understanding. This chapter delves into panoramic semantic segmentation techniques for achieving omnidirectional scene understanding. To learn from the label-rich pinhole image domain, we for the first time, propose a new setting of unsupervised domain adaptation (UDA) for panoramic semantic segmentation. A new dataset is created for benchmarking panoramic semantic segmentation. Based on this UDA setting, a novel domain adaptation framework is first proposed to address the transfer learning for panoramic images. Furthermore, a distortion-aware transformer model is designed to improve the segmentation performance before and after domain adaptation. These methodologies are employed to achieve comprehensive scene interpretation within the context of ITS.

Chapter 4: Towards Robust Scene Understanding. Focusing on multimodal semantic segmentation, this chapter explores the fusion methods of various sensory data sources, contributing to robust scene understanding for intelligent mobility. A unified transformer model is proposed to handle fusion of diverse modalities for semantic segmentation. A new benchmark is constructed for RGB-Event semantic segmentation. Apart from bi-modal fusion, an advanced unified model is proposed to deal with arbitrary-modal semantic segmentation, which can combine up to 80 modalities with RGB. Besides, a new dataset for arbitrary-modal semantic segmentation is presented. These approaches are investigated to maintain robust scene understanding for ITS.

PART III: MOBILITY ASSISTANCE SYSTEMS (MAS)

Chapter 5: Towards Navigational Scene Understanding. This chapter is dedicated to advancing the capabilities of Mobility Assistance Systems (MAS) through two domains, each aimed at empowering visual localization and fostering semantic mapping for enhanced accessibility. In the pursuit of robust visual localization, the first research theme introduces a novel transformer model with interleaving attention for feature matching. This innovative approach is tailored to both indoor and outdoor visual localization scenarios, mitigating the challenges posed by texture-less scenes and enabling robust and precise visual localization. The second research theme delves into the field of semantic mapping, a critical component in enabling autonomous navigation for individuals with diverse mobility needs. A cutting-edge model, Trans4Map, is introduced to facilitate end-to-end training for semantic mapping, and amplify the potential for independent and safe navigation. The third research theme is to achieve holistic semantic mapping via a single panoramic image. This new task is named as 360BEV. Two datasets are extended to enable the end-to-end training for panoramic semantic mapping. Through these research contributions, this chapter establishes a comprehensive framework for navigational perception of Mobility Assistance Systems.

Chapter 6: Towards Generalizable Scene Understanding. This chapter addresses the practical concern in ITS and MAS. For example, the challenge posed by transparent objects in real-world scenarios, can be particularly daunting for people with visual impairments. Here, we present a novel solution to transparent object segmentation. This pioneering approach seeks to enhance the safety and accessibility of navigation for both blind and sighted pedestrians. Particularly, this approach serves as the foundation of our wearable MAS, called *Vision4Blind*. Furthermore, we extend the dual-head vision transformer model to encompass the adverse driving perspective. Moreover, a novel Multi-source Meta-learning UDA (MMUDA) framework is proposed to transform models from normal to abnormal (accident) scene segmentation.

Chapter 7: Assistive Systems and Applications. This chapter introduces our iterative development on the *Vision4Blind* system. It combines smart glasses equipped with cameras and a portable GPU processor. The efficacy of this system is assessed through a series of field tests and user studies. Besides, we conduct a proof-of-concept exploration of novel Mobility Assistance Systems (MAS). For example, we explore the utilization of drones as aides for the navigation of People with Visual Impairments (PVI), *i.e.*, as a “flying guide dog”. The application not only empowers the drone to semantic segmentation but also paves the way for the recognition of pedestrian and vehicle traffic lights. This pioneering concept is validated through a comprehensive user study involving blindfolded participants, providing empirical evidence of the system’s potential to enhance navigation.

PART IV: INSIGHTS OF SCENE UNDERSTANDING

Chapter 8: Contributions. This chapter concludes the contributions made in ITS and MAS. The contribution of dataset lies in the creation of two new datasets. The first one, tailored for panoramic semantic segmentation, enables evaluation of models for omnidirectional scene understanding. The second dataset, centered on arbitrary-modal semantic segmentation, extends the frontiers of knowledge by facilitating the fusion of diverse sensory inputs for enhanced understanding. Besides, novel methodologies have been forged to empower both ITS and MAS. The novel methods include omnidirectional and multimodal segmentation, transparent object segmentation, visual localization, semantic mapping, and more. Besides, three new systems are implemented to assist the people with visual impairments. These human-friendly systems are iteratively implemented and enhanced by considering feedback from target users.

Chapter 9: Outlook. This chapter engages in a discussion about the next-generation assistance systems and also the future work of scene understanding. Some promising research directions related to the scene understanding for ITS and MAS are presented in this chapter. One important research topic is the vision-language navigation for helping people with visual impairments. Besides, using one unified model to address all vision-based even vision-language tasks is another future work to implement the next-generation assistance system. The Artificial General Intelligence (AGI) also has the potential to foster development of assistive technology even further.

2

RELATED WORK

> The central objective of this thesis is to explore the empowerment of scene understanding to create a smart mobility utopia. We focus on the image segmentation task and delve into two fields: (1) intelligent transportation systems and (2) mobility assistance systems. This chapter presents an overview of the most relevant literature according to scene understanding, covering tasks from traditional to novel semantic segmentation tasks, a detailed study of benchmarks according to different focuses, and various practical assistance systems.

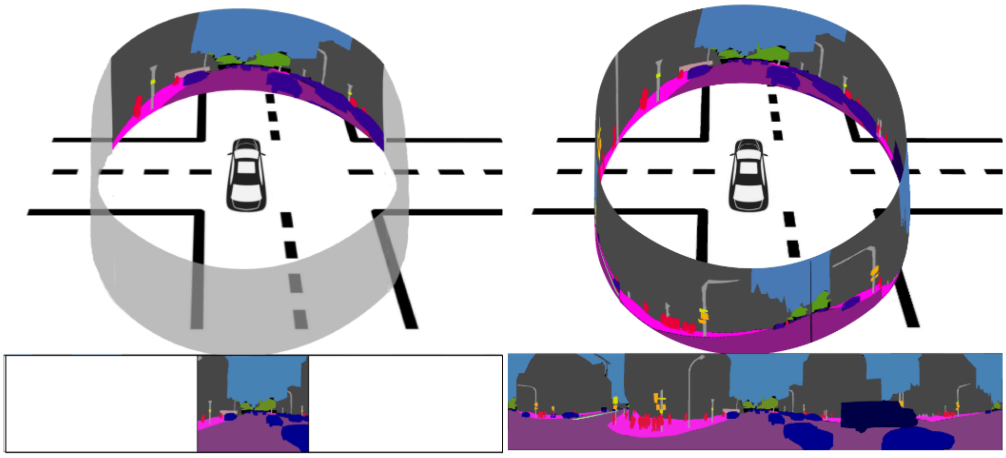
2.1 SCENE UNDERSTANDING THROUGH IMAGE SEGMENTATION

Scene understanding stands as a complex challenge within the field of computer vision, with its primary objective being the extraction of meaningful insights from images. This crucial information is derived through diverse segmentation settings.

2.1.1 *Image Semantic Segmentation*

The task of image semantic segmentation entails the partitioning of images into distinct regions, each associated with semantically meaningful categories, such as *car* and *road* from outdoor scenes, *table* and *chair* from indoor scenes, etc. This finer granularity allows for the discernment of object boundaries and the creation of detailed scene interpretations. Dense image semantic segmentation has garnered significant attention and witnessed remarkable advancements since the inception of Fully Convolutional Networks (FCN) [127], which introduced it as an end-to-end per-pixel classification task. Building on the foundation laid by FCN, subsequent endeavors have pushed the boundaries of segmentation performance by embracing encoder-decoder architectures [29, 208], amplifying the potential of high-resolution representations [116, 217], expanding receptive fields [28, 76, 293], and collecting contextual priors [90, 274].

Drawing inspiration from non-local blocks [223], the integration of self-attention mechanisms [213] has proven instrumental in establishing long-range dependencies [61, 85, 110, 124, 263] within the framework of FCNs. Modern architectural innovations have even led to the substitution of conventional convolutional backbones with transformer architectures [52, 207], ushering in a perspective where image understanding can be envisaged as a form of sequence-to-sequence learning. This paradigm shift is evident through the rise of dense prediction transformers [51, 108, 125, 222, 229] and semantic segmentation transformers [39, 69, 191, 238, 296], encapsulating the essence of image interpretation within a transformer-based framework. Adding to this evolution, recent explorations have introduced architectures akin to Multilayer Per-



(a) Semantic segmentation in narrow FoV (b) Semantic segmentation in omni FoV

Figure 3: Comparison of pinhole and panoramic semantic segmentation in driving scenarios with different field of view (FoV).

ceptrons (MLPs) [34, 75, 112, 206], which ingeniously alternate between spatial- and channel-based mixing. This novel approach has sparked substantial interest, offering a promising avenue for tackling a wide spectrum of visual recognition tasks.

However, most of these methods are tailored to narrow Field-of-View (FoV) pinhole images and often tend to exhibit notable accuracy degradation when extended to the 360° domain for holistic scene understanding. As shown in Figure 3, the ultra-wide FoV from 360° cameras might bring image distortion and object deformation.

2.1.2 *Panoramic Semantic Segmentation*

In the pursuit of holistic scene understanding, panoramic semantic segmentation emerges as a vital consideration. This task involves semantically segmenting 360° images, providing an omnidirectional view that is especially valuable in applications like autonomous driving. On the one hand, outdoor omnidirectional semantic segmentation systems rely on fisheye cameras [49, 152, 257] or panoramic images [81, 145, 244].On the other hand, indoor methods focus on either distortion-mitigated representations [86, 99, 177, 203, 297] or multi-tasks schemes [122, 192, 270]. A common underlying assumption in these methods is the availability of labeled images, either fully or partially, in the panorama domain for training segmentation models.

However, obtaining dense pixel-wise labels is an exceedingly labor-intensive and time-consuming endeavor, especially for panoramas featuring higher complexities and an abundance of small objects present in wide-FoV observations. In order to alleviate the need for labeled target data and mitigate the prohibitively expensive annotation process required for determining pixel-level semantics within unstructured real-world environments, we propose an innovative approach. Specially, we look into panoramic semantic segmentation via the lens of unsupervised transfer learning. We explore the Pinhole-to-Panoramic (Pin2Pan) adaptation strategy, capitalizing on the potential of rich and readily available datasets, *e.g.*, annotated pinhole datasets.

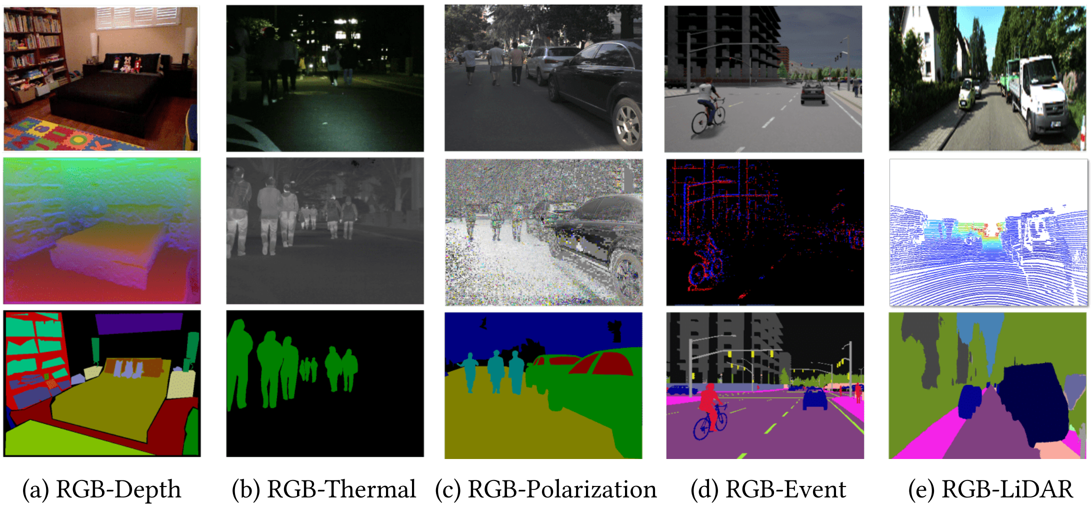

(a) RGB-Depth (b) RGB-Thermal (c) RGB-Polarization (d) RGB-Event (e) RGB-LiDAR

Figure 4: Different multimodal semantic segmentation. From top to bottom are the RGB image, the X modality, and the ground-truth segmentation.

2.1.3 *Multimodal Semantic Segmentation*

Addressing the growing complexity of real-world scenes, multimodal semantic segmentation tackles the fusion of diverse information. Approaches like RGB-Depth, -Thermal, -Polarization, -Event, -LiDAR fusion (Figure 4) augment the capabilities of segmentation models, resulting in more accurate and robust scene understanding.

Approaches such as RGB-Depth [158, 301] and RGB-Thermal [196, 198, 305] semantic segmentation have garnered significant attention. Moreover, the utilization of polarimetric optical cues [95] and event-driven priors [286] has been explored to ensure reliable perception under adverse conditions. In the context of automated driving, the inclusion of LiDAR data [309] has enhanced semantic understanding of the road scene. Nevertheless, most of these studies focus on a single modality combination.

In multimodal semantic segmentation, two predominant strategies have emerged. The first approach involves incorporating cross-modal complementary information into layer- or operator-based designs [20, 31, 220, 233, 241]. While these studies demonstrate the acquisition of multimodal features within a shared network, their designs are often tailored for specific modalities, such as RGB-D semantic segmentation, which limits their adaptability to other modalities. Furthermore, certain multi-task frameworks [7, 291] facilitate inter-task feature propagation for RGB-D scene understanding, but they depend on supervision from other tasks for joint learning. The second paradigm focuses on devising fusion schemes to bridge parallel modality streams. For instance, ACNet [82] introduces attention modules to exploit features for RGB-D semantic segmentation, while ABMDRNet [287] aims to reduce modality differences in features before extracting discriminative cues for RGB-T fusion. In RGB-P segmentation, Xiang *et al.* [235] connected RGB and polarization branches through channel attention. For RGB-E parsing, Zhang *et al.* [285] explored sparse-to-dense and dense-to-sparse fusion flows to extract dynamic context for accident scene segmentation.

Towards robust scene understanding, we tackle RGB-X semantic segmentation within a unified framework, enabling diverse modality combinations.

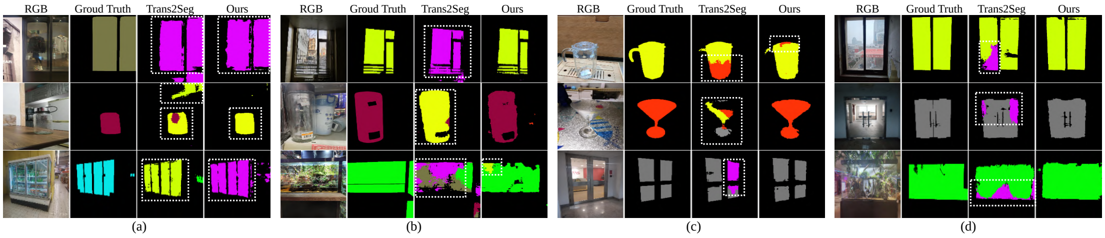


Figure 5: Examples of transparent object segmentation on Trans10K-v2 test set.

2.1.4 *Transparent Semantic Segmentation*

Transparent objects present distinct challenges in navigational scene understanding, particularly in hazard and edge cases. Transparent semantic segmentation, a new research direction, aims to identify and delineate glass-like objects, contributing to navigational perception and mobility, especially for pedestrians and individuals with visual impairments. Some examples are presented in Figure 5.

Traditional visual assistance systems [8, 84] have employed multi-sensor fusion to tackle challenges associated with transparent obstacles, such as glass objects, French windows, and French doors. These systems often combined ultrasonic sensors with RGB-D cameras to effectively address such scenarios. Furthermore, they frequently harnessed multimodal and multispectral information for enhanced perception. Okazawa *et al.* [144] achieved simultaneous recognition of both conventional non-transparent objects and transparent objects by leveraging the differences in transmission characteristics within multispectral scenes. Additionally, polarization cues [95] and reflection priors [117] have been extensively investigated for transparency perception. Notably, Xiang *et al.* [235] developed a polarization-driven semantic segmentation architecture that dynamically bridges RGB and polarization dimensions using efficient attention connections. This approach leverages the optical features of polarimetric information to robustly represent diverse materials, enhancing the segmentation performance of classes with polarization-specific properties, such as *glass*.

2.2 BENCHMARKS FOR SEMANTIC SEGMENTATION

This section dives into the existing benchmarks that facilitate rigorous evaluation of semantic segmentation algorithms, including the perspective of driving scenes, general scenes, corner cases, and multimodal fusion.

2.2.1 *Focusing on Driving Scenes*

Semantic segmentation datasets dedicated to driving scenes offer a valuable testbed for models developed for transportation applications. These datasets encompass a wide array of driving scenarios, reflecting the challenges faced by intelligent transportation systems. While numerous semantic segmentation datasets offer valuable training and testing data, we only highlight a few of the most relevant ones in the following.

Cityscapes. The Cityscapes [44] dataset is one of the most widely used benchmarks for semantic segmentation in driving scenarios. It has 5,000 high-resolution images captured from 50 different cities in Europe, and includes 19 classes.
KITTI-360. The KITTI-360 [114] is a suburban driving dataset, having 49,004/12,276 images at the size of 1408×376 for training/validation with 19 classes.
BDD100K. The BDD100K [260] dataset provides a large-scale collection of images captured from diverse driving scenarios. A keyframe at the 10th second from each sequence has dense annotation with 19 semantic classes.
IDD. The IDD [211] dataset is captured from 182 drive sequences in Indian cities. It consists of 10,000 images annotated with 34 classes.
Mapillary Vistas. The Mapillary Vistas [142] dataset focuses on capturing street-level imagery from different parts of the world, providing 25,000 high-resolution images annotated into 66 object categories with additional, instance-specific labels for 37 classes.
ApolloScape. The ApolloScape [83] dataset is tailored for autonomous driving research. It includes 143,906 video frames and corresponding pixel-level annotations. The images are labeled with 25 semantic classes.

Compared to these datasets, in the pursuit of holistic scene understanding, we propose a new dataset with 360° images, namely *DensePASS*, which covers 100 labelled data with 19 classes and 2,000 unlabelled data.

2.2.2 *Focusing on General Scenes*

Beyond specific driving contexts, benchmarks focusing on general scenes provide a broader perspective on the capabilities of semantic segmentation models. These datasets encompass a diverse range of scenes and scenarios as well.
ADE20K [299] is a large-scale dataset for semantic segmentation of indoor and outdoor scenes. It contains 20,288 images with 150 semantic classes, including both common and rare objects. The images are of high quality and diversity.
COCO-stuff [17] is a subset of the COCO dataset that is specifically designed for semantic segmentation of stuff classes. Stuff classes are classes that are not considered to be objects, such as sky, ground, and vegetation. COCO-stuff contains 171 stuff classes.

2.2.3 *Focusing on Corner Cases*

Addressing challenging and adverse scenarios, benchmarks centered around corner cases test the limits of semantic segmentation models. This includes transparent object segmentation, adverse driving scenes, and accidental semantic segmentation, helping to push the boundaries of what these models can accurately interpret.
WildDash. The WildDash [266] is a dataset of challenging driving scenarios for testing the robustness of semantic segmentation. WildDashv2 [267] is an extended version, which contains 5,032 images with 26 classes in evaluation version.
ACDC. The ACDC [171] dataset has adverse driving conditions for semantic segmentation. It contains 4,006 images with 19 semantic classes. The images are collected in a variety of adverse conditions, such as rain, snow, and fog. The dataset is designed to test the robustness of semantic segmentation models to adverse weather conditions.

DADA-seg. The DADA-seg [285] is a dataset of accidental semantic segmentation for autonomous driving. It contains 313 images with 19 semantic classes for testing. The images are collected in a variety of scenarios where accidental semantic segmentation can occur, such as when a vehicle is partially obscured by another vehicle or object.
Trans10K-v2. The Trans10K-v2 [240] is a dataset for RGB-based transparent object segmentation. There are 11 categories marked as *shelf, jar or tank, freezer, window, glass door, eyeglass, cup, wall, glass bow, water bottle*, and *storage box*.

2.2.4 *Focusing on Multimodal Fusion*

Various multimodal semantic segmentation benchmarks contribute to the advancement of multimodal fusion techniques, which leverage the synergistic capabilities of different sensors. In this following, we list the most relevant multimodal datasets.
KITTI-360 [114] dataset can also be used for RGB-Depth-Event-LiDAR fusion, after the depth images and event data are generated. There are 19 semantic classes.
MFNet [71] is an urban street dataset with 1,569 RGB-Thermal pairs at the size of 640×480 with 8 classes. 820 pairs are captured from day-time scenes and 749 are captured from night-time scenes.
NYU Depth V2 [185] is an indoor understanding dataset with 1,449 RGB-Depth pairs at the size of 640×480, splitting into 795/654 for training/testing with 40 classes.
SUN-RGBD [188] dataset has 10,335 RGB-Depth images with 37 classes, and 5285/5050 for training/testing, respectively.
Stanford2D3D [4] dataset has 70,496 RGB-Depth images with 13 object categories. Areas of {1, 2, 3, 4, 6} are used for training and area 5 is for testing.
ScanNetV2 [45] dataset provides 19,466/5,436/2,135 RGB-Depth samples for training/-validation/testing. There are 20 classes.
RGB-P ZJU [235] is an RGB-Polarization dataset collected by a multimodal vision sensor designed for automated driving on complex campus street scenes. It is composed of 344 images for training and 50 images for evaluation, both labeled with 8 semantic classes. The input image is resized to 612×512.
UrbanLF [180] is a light field dataset with real and synthetic sets annotated in 14 classes, respectively splitting into 580/80/164 and 172/28/50 samples for training/validation/testing. Each sample is composed of 81 sub-aperture images.
MCubeS [113] is a dataset with pairs of RGB, Near-Infrared (NIR), Degree of Linear Polarization (DoLP), and Angle of Linear Polarization (AoLP), to study semantic material segmentation of 20 classes. It has 302/96/102 image pairs for training/validation/testing at the size of 1224×1024.

To provide a diverse multimodal semantic segmentation benchmark, we spent the effort to create a large-scale dataset DeLiVER with Depth, LiDAR, Views, Event, RGB data, based on the CARLA simulator [53]. DeLiVER provides six mutually orthogonal views (*i.e.*, *front, rear, left, right, up, down*) of the same spatial viewpoint, *i.e.*, a complete frame of data is encoded in the format of a panoramic cubemap. Besides, it features severe weather conditions and five sensor failure modes to exploit complementary modalities and resolve partial sensor outages in realistic driving scenarios.

2.3 VISUAL ASSISTANCE SYSTEMS

As technology advances, vision-based assistance systems play a pivotal role in enhancing accessibility and independence for individuals with visual disabilities. Visual assistance systems are used to provide essential environmental information to achieve the navigation of the visually impaired through wearable sensors [38, 102, 120] and environment perceiving sensors [54, 121, 205, 281].

2.3.1 *Scene Segmentation*

Applying advanced semantic segmentation methods in assistive technology can empower users by providing human-friendly information about their surroundings, aiding in navigation and scene understanding. Semantic segmentation is a well-established technique used to assist the visually impaired. It involves assigning a semantic label to each pixel in an image, such as *road*, *sidewalk*, or *bicycle*. This information can be used to help visually impaired people navigate their surroundings safely and independently.

Previous works [54, 105, 106, 121, 205, 247, 295, 310] explore semantic segmentation to create assistance systems for visually impaired people by providing a deep understanding of the scenario to achieve better assistance. Yang *et al.* [247] unified intersection-centered perception tasks by utilizing real-time semantic segmentation. In [121], a lightweight system with a solid-state LiDAR sensor is proposed for holistic indoor detection and obstacle avoidance using 3D point cloud instance segmentation. The system implements obstacle avoidance and object finding together with voice guidance, so that the user can scan new point clouds from a changing indoor environment. Tian *et al.* [205] and Li *et al.* [106] concentrated on handling the crosswalk situation, which involves the segmentation of objects and prediction of the traffic light status, while Zou *et al.* [310] focused on real-time passable area segmentation. Most of the above works rely on Convolutional Neural Networks (CNN) as feature extractors. Ma *et al.* [135] proposed a robot system to achieve a wayfinding function for the blind. Zheng *et al.* [295] focused on material recognition in wearable robotics.

2.3.2 *Obstacle Avoidance*

Obstacle avoidance systems have been explored and implemented to assist the safe navigation through complex environments in indoor and outdoor spaces. Rodriguez *et al.* [163] segmented the image into background and obstacles based on dense disparity maps and ground plane estimation algorithms. Sonar-based obstacle avoidance systems [159] use sonar sensors to detect obstacles in the environment, which is relatively inexpensive and easy to implement but can be less accurate. In [186], an indoor navigation wearable system based on visual markers recognition and ultrasonic obstacles perception is proposed and utilized as an audio assistance for people with visual impairments. Vision-based obstacle avoidance systems [9] use cameras to detect obstacles in the environment. The camera captures images of the environment and then uses computer vision algorithms to identify and track obstacles.

2.3.3 *Orientation and Navigation*

Vision-based approaches contribute to various components of orientation and navigation, from accurate visual localization [259, 268, 276] to the creation of semantic maps [19, 22, 57, 67, 111, 149, 169] that assist users in understanding and navigating their surroundings effectively.

A visual positioning system [276] uses an RGB-D camera and an inertial measurement unit to estimate the pose and utilizes depth-enhanced visual-inertial odometry for indoor navigation. DS-SLAM [259] combines semantic segmentation network with a moving consistency check method to reduce the impact of dynamic objects, thus significantly improving localization accuracy in dynamic environments. A frame-to-frame VO algorithm [268] combines deep learning with epipolar geometry and Perspective-n-Point method by training two convolutional networks.

In recent times, a multitude of approaches have arisen in the field of semantic mapping. Semantic SLAM pipelines [68, 160] are instrumental in constructing such maps. These pipelines involve forwarding images into segmentation networks and subsequently projecting predicted labels onto top-view maps. The project-then-segment pipeline [187] can lead to significant loss of visual information during projection, particularly hampering small object segmentation. SMNet [22] follows an offline project-then-segment methodology, training encoder and decoder networks through two stages. Lu *et al.* [129] introduced an end-to-end network that encodes front-view information of the driving scene, subsequently decoding it into a 2D top-down view. Pan *et al.* [149] presented a cross-view network incorporating a view parsing network to parse semantics across diverse views. Moreover, there are many Bird's-Eye-view (BEV) semantic segmentation approaches for driving scene perception [19, 57, 67, 111, 169] emerging in the field. BEVFormer [111] aggregates spatio-temporal cues from surround-view cameras, whereas ViT-BEVSeg [56] uses a spatial transformer decoder for generating semantic occupancy grid maps.

Apart from these mapping methods, a wearable navigation system for visually impaired and blind people in unknown indoor and outdoor environments was proposed in [93]. The system can map and track the position of the pedestrian during exploration of the new environment. Lin *et al.* [120] proposed a learning-based wearable system to achieve the navigation for visually impaired. V-Eye [54] made use of a global localization method to pursue a better scene understanding, while the outdoor walking guide system [77] leverages depth information. Cao *et al.* [21] developed a light-weight network for fast detection of blind road and sidewalk. To address the challenges posed by the COVID-19 pandemic, an object-finding algorithm is introduced in [1] to build an end-to-end perception robotic cane system, which can enable socially-preferred autonomous goal selection and navigation in indoor spaces. The work of "*I am the follower, also the boss*" [289] uses machine forms of a guiding robot and anatomy from different stages to achieve visually impaired assistance.

As one of the two major research fields in this thesis, we delve deep into developing Mobility Assistance Systems (MAS) for assisting pedestrians as well as helping people with visual impairments. The research theme of MAS includes navigational scene understanding, realistic scene understanding and more novel concepts and prototypes.

Part II

INTELLIGENT TRANSPORTATION SYSTEMS

3 TOWARDS OMNIDIRECTIONAL SCENE UNDERSTANDING

> This chapter presents the first research theme in the field of ITS, *i.e.*, *omnidirectional scene understanding* through panoramic semantic segmentation. The challenges include the limited availability of annotated panoramic data and the distortion of panoramic images. To address that, our contributions are two-fold: (1) We pioneer a novel approach, *Pinhole-to-Panoramic Domain Adaptation (P2PDA)*, which constitutes a new Unsupervised Domain Adaptation (UDA) setting to bridge the gap between pinhole and panoramic image domains. This P2PDA framework is presented in Section 3.1, based on our work published in *Transactions on ITS* 2022 [280]. (2) To address image distortions, we propose distortion-aware vision transformer models for panoramic segmentation, *i.e.*, *Trans4PASS*, which is detailed in Section 3.2, based on our *CVPR 2022* publication [283].

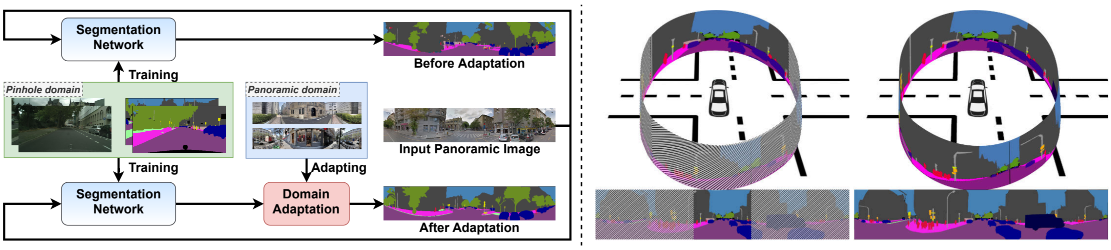
3.1 PINHOLE-TO-PANORAMIC DOMAIN ADAPTATION

This section is based on our work published in *Transactions on ITS 2022* [280].

3.1.1 *Motivation of P2PDA*

Semantic segmentation accuracy has increased at a rapid pace thanks to the resilience of advanced neural networks. However, most of the previous frameworks were developed with the assumption that the driving scene images are captured with a *pinhole* camera [44, 65, 114], which has a comparably narrow Field of View (FoV). This limits the capabilities of scene perception systems. While mounting multiple sensors can mitigate this issue, it requires additional data fusion and sensor calibration [11, 49]. Recently, a novel approach for expanding FoV [248, 249] has emerged: using a single *panoramic* camera to perform 360° scene understanding, as shown in Figure 6b.

However, the lack of pixel-wise annotations for panoramic images is a major obstacle to the advancement of semantic segmentation research for this type of data. However, recent progress in the field of domain adaptation (DA) has led to the development of highly effective techniques that can be used to complement the limitation of training data in driving scenarios, such as nighttime driving [165] and accident scenes [285]. In this work, we address the challenge of label-scarce panoramic segmentation by adopting DA, transferring knowledge from considerably larger datasets of the pinhole im-



(a) Pinhole-to-Panoramic Domain Adaptation (b) Narrow and 360° Field of View (FoV)

Figure 6: An overview of the formalized task of domain adaptation for panoramic semantic segmentation. The *source* domain (green) contains *pinhole* images with annotations, while the *target* domain (blue) contains *panoramic* images without annotations. (b) FoV comparison between pinhole forward-view and 360° panoramic surround-view imaging of self-driving scenes.

age domain. The problem statement is structured as: *Unsupervised Domain Adaptation (UDA) in panoramic semantic segmentation*, which involves the adaptation from the label-rich pinhole (*source*) domain to the label-scarce panoramic (*target*) domain. The overview of the formalized task is shown in Figure 6a.

To promote research on panoramic semantic segmentation under cross-domain conditions, we introduce a new dataset – *Dense PAnoramic Semantic Segmentation (DensePASS)* – covering 360° images captured from all over the world to ensure diversity. Our benchmark provides (1) an unlabelled panoramic training set for optimizing the domain adaptation model and (2) a panoramic test set manually labelled with 19 classes following Cityscapes [44], a dataset of pinhole images that we use as the label-rich training data from the source domain.

According to our observations, directly transferring models trained on pinhole images to panoramic data often results in a significant drop in accuracy. This is because panoramic images have a different layout than pinhole images, due to the equirectangular projection. As shown in Figure 6b, panoramic images have a longer horizontal distribution and geometric distortion on both sides of the viewing direction. This results in a considerable domain shift, which can significantly degrade the performance of the model. To address the challenge of label-scarce panoramic segmentation, we implement *P2PDA*, a generic framework for *Pinhole to Panoramic Domain Adaptation*.

3.1.2 *The new DensePASS Dataset*

There is a lack of established segmentation benchmarks that address the challenging task of Pinhole→Panoramic recognition. Additionally, previous panoramic testbeds only cover a very limited number of classes [249, 252]. To address these limitations, we collect DensePASS – a novel densely annotated dataset for panoramic segmentation of driving scenes. DensePASS is created with the Pinhole-to-Panoramic transfer in mind, and the test data is annotated with 19 categories that are also present in the pinhole camera dataset Cityscapes [44] and other prominent semantic segmentation benchmarks [260, 266]. To facilitate the unsupervised domain adaptation task, DensePASS covers both, labelled data (100 panoramic images used for testing) and unlabelled training data (2000 panoramic images used for the domain transfer optimization). A FoV of 70°×360° is covered in the captured panoramic images with a 400×2048 resolu-

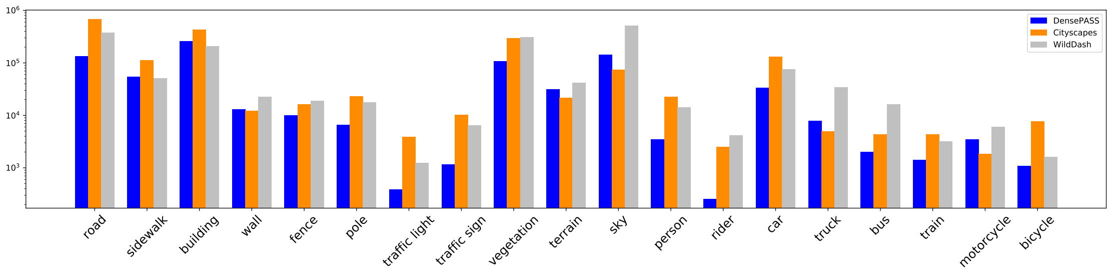


Figure 7: Distribution of the DensePASS dataset in terms of class-wise pixel counts per image.

tion. The data is collected using Google Street View and includes images from different continents (25 different cities for testing and 40 for training).

In Figure 7, we compare label distributions of DensePASS with Cityscapes [44] and WildDash [266] datasets in terms of pixel counts by averaging the number of images used in our domain adaptation study. Our histogram analysis indicates, that DensePASS and the mentioned pinhole camera datasets follow a relatively close distribution of categories. This observation indicates, that distribution alignment not only in the feature-space but also in the semantic output-space might be beneficial and is therefore integrated in our framework. Overall, DensePASS is a valuable new resource for the task of panoramic segmentation of driving scenes. It provides a large and diverse dataset of labelled and unlabelled images, and it is carefully designed to address the challenges of Pinhole-to-Panoramic UDA. We believe that DensePASS will be a valuable tool for researchers and developers working on this important problem.

3.1.3 *P2PDA Framework*

3.1.3.1 *Framework Overview*

Attention-augmented adversarial adaptation. Compared to AdaptSegNet architecture [208], our framework has multiple variants of region- or attention-augmented DA modules plugged in at different network depths with an overview provided in Figure 8. The main components of our framework are a weight-shared segmentation network G with attention modules and multiple DA modules equipped with the corresponding discriminators D. For unsupervised domain adaptation methods, only the source domain dataset $\mathcal{D}_s = \{(x_s, y_s)|x_s \in \mathbb{R}^{H_s \times W_s \times 3}, y_s \in \mathbb{R}^{H_s \times W_s \times 1}\}$ and the unlabelled target domain dataset $\mathcal{D}_t = \{(x_t)|x_t \in \mathbb{R}^{H_t \times W_t \times 3}\}$ are given, where x_s and x_t denote the input images from source and target domains, and y_s are the ground truth labels in source domain. We note that (H_s, W_s) and (H_t, W_t) are the height and width of the source and target images, respectively.

For ease of understanding, we only list the formulas for a single classifier on G and single D. For multiple G or D, they will be combined through specific hyperparameters. First, the source domain images x_s are fed into the segmentation network G (also referred to as the generator) to create prediction results $\tilde{y}_s$=$G(x_s)$ and the source ground-truth labels y_s are used to compute the segmentation loss $\mathcal{L}_{seg}$:

$$\mathcal{L}_{seg}(G) = \mathbb{E}\left[\ell(G(x_s), y_s)\right], \tag{1}$$

where $E[\cdot]$ is the statistical expectation and $\ell(\cdot,\cdot)$ is the standard cross entropy loss.

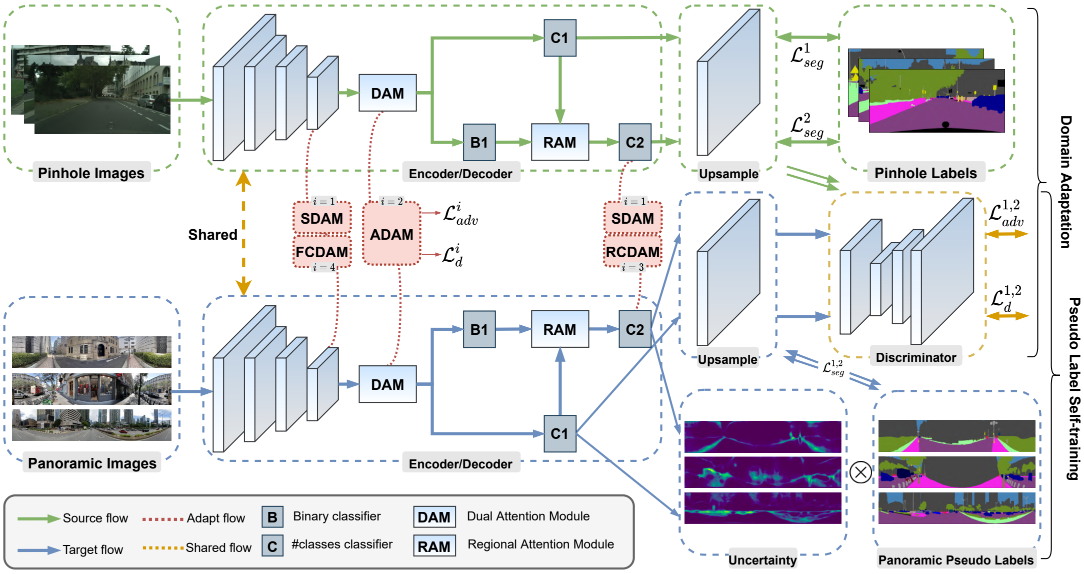


Figure 8: Diagram of the P2PDA framework. The shared backbone is an encoder-decoder segmentation network (*e.g.*, DANet). The SDAM module is applied on the high-level feature-space or output-space and the FCDAM on the feature confidence space, while ADAM is performed on the dual attended feature-space and RCDAM on the output-space after the region attention module. In the second stage of pseudo-label self-supervised learning, the uncertainty map is calculated based on C1 and C2 predictions and used for element-wise multiplication with pseudo-labels as an online selection of pseudo-labels.

Next, the discriminator D is trained with the binary objective to distinguish between the source and target domains of the input, so the discriminator loss is formulated as:

$$\mathcal{L}_d(D) = \mathbb{E}\left[\ell(D(G(x_s)), 0)\right] + \mathbb{E}\left[\ell(D(G(x_t)), 1)\right], \tag{2}$$

where $\ell(\cdot, \cdot)$ is the binary cross entropy, with 0 and 1 being the two-class labels (pinhole and panoramic).

Then, to enforce the generator G to align the distribution of $\tilde{y}_t$ closer to $\tilde{y}_s$, the prediction results $\tilde{y}_t = G(x_t)$ for the target domain is directly used to estimate the adversarial loss, which is updated alongside with $\mathcal{L}_{seg}$ and is formulated as:

$$\mathcal{L}_{adv}(G) = \mathbb{E}\left[\ell(D(G(x_t)), 0)\right]. \tag{3}$$

The adversarial loss is high if the discriminator prediction is correct. This means that the adversarial loss encourages the segmentation network to generate segmentation masks in the target domain that are indistinguishable from the masks in the source domain. In other words, the discriminators are trained to distinguish between the source and target domains with $\mathcal{L}_d(D)$, while the segmentation network G is trained to (1) correctly segment the images from the source domain with $\mathcal{L}_{seg}$, and (2) make the target domain data indistinguishable from the source data by fooling the discriminator.

The join loss from Eq. (1) and Eq. (3) used to train the generator G becomes:

$$\mathcal{L}(G) = \lambda_{seg}\mathcal{L}_{seg}(G) + \lambda_{adv}\mathcal{L}_{adv}(G), \tag{4}$$

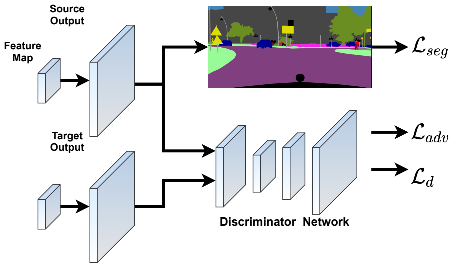


Figure 9: The segmentation domain adaptation module (SDAM).

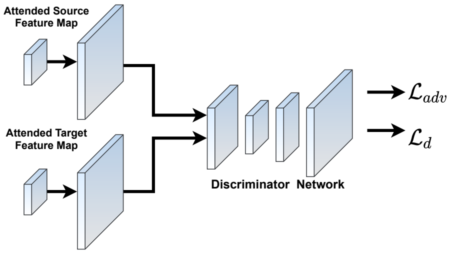


Figure 10: The attentional domain adaptation module (ADAM).

where λ_{adv} and λ_{seg} are weights used to balance the domain adaptation and semantic segmentation losses. To perform end-to-end training for multiple classifiers of G and multiple D, our final loss function is denoted as:

$$\mathcal{L}(G, D) = \sum_i \lambda^i_{seg}\mathcal{L}^i_{seg}(G) + \sum_i \lambda^i_{adv}\mathcal{L}^i_{adv}(G) + \sum_i \lambda^i_{d}\mathcal{L}^i_{d}(D). \tag{5}$$

Attention-regulated self-learning adaptation. Our network can readily generate highly qualified segmentation masks on panoramic images, after the main stage of multi-level alignment through the domain adaptation modules described in Section 3.1.3.2. Our next goal is to advance the training procedure by using the inherent knowledge in the pixel-wise predictions from the first training stage, *i.e.*, *panoramic pseudo-labels*. To achieve this, P2PDA uses an uncertainty-aware domain adaptation stage to improve the prediction in an iterative fashion. In this stage, the source images are replaced by the self-supervised panoramic images, *i.e.*, the predictions are used to refine the model itself. The key idea of this training stage is to employ multiple classifiers with attention heads naturally encouraged to produce discrepant predictions in order to assess the uncertainty of the pseudo-labels. First, we estimate the uncertainty map by using the variance operation on predictions produced with two different classifiers with disparate attention modules as in [61, 178]. Then, we apply element-wise multiplication of the pseudo-labels with the resulting uncertainty map and, finally, we threshold the resulting value to obtain the certain pseudo-labels. An overview of the uncertainty-driven self-training is shown in the bottom part of Figure 8.

3.1.3.2 *Domain Adaptation Modules*

Segmentation domain adaptation module. As illustrated in Figure 9, our initial module SDAM is derived from AdaptSegNet and attempts to match the source and target segmentation outputs. After a segmentation network forward pass with images from both domains (x_s and x_t), feature maps of both representations are used as input to the discriminator D which learns to distinguish the domain with $\mathcal{L}_d(D)$, while the segmentation network G learns to segment the pinhole images with $\mathcal{L}_{seg}(G)$ and align the domains with $\mathcal{L}_{adv}(G)$. SDAM learns a Pinhole-to-Panoramic domain adaptation model at multiple levels jointly within our P2PDA framework, as shown in Figure 8.

Attentional domain adaptation module. Next, we design *ADAM*, an *attentional* domain adaptation module, aimed at detecting and magnifying the significant amount of pinhole-panoramic correspondences at both, local and global levels (overview in Figure 10). ADAM differs from SDAM as it leverages the attention mechanism to learn an

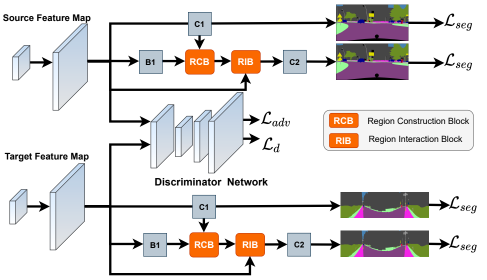


Figure 11: The regional context domain adaptation module (RCDAM).

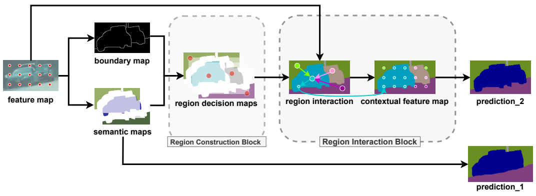


Figure 12: Diagram of Region Construction and Interaction Blocks.

optimal weighting scheme for the discriminator input. As in the Dual Attention Module (DAM) [61] shown in Figure 8, the feature map extracted by the backbone model is denoted as $F \in \mathbb{R}^{h \times w \times c}$, where the h, w, and c are the height, width, and channel of the feature map. After this representation is reshaped as $F' \in \mathbb{R}^{(h \times w) \times c}$, the position-wise attended feature is calculated as: $S = \sigma(F'^T, F' \otimes F'^T), S \in \mathbb{R}^{c \times (h \times w)}$, where σ is the Softmax function. Similarly, the channel-wise attended feature is denoted as $R = \sigma(F', F'^T \otimes F'), R \in \mathbb{R}^{(h \times w) \times c}$. Then, the final dual attended feature is concatenated with the the reshaped $S' \in \mathbb{R}^{h \times w \times c}$ and the reshaped $R' \in \mathbb{R}^{h \times w \times c}$. By doing this, ADAM enables direct context information exchange among all pixels, mitigating the influence of discrepancy in positional priors and local distortions. Relevant portions of the feature maps of both, x_s and x_t inputs are enhanced through the attention and the *re-weighted* source and target representations are both used to optimize the corresponding discriminator D.

Regional domain adaptation module. Next, we focus on *region relationship of the panoramic images*. Inspired by RANet, we design the RCDAM module based on the Regional Attention Module (RAM) [178] to configure the information flow between different regions and within the same region, as illustrated in Figure 11. RCDAM follows a hierarchical adversarial learning scheme with two-stage discriminators, where the first stage is identical to the previously described SDAM. The second stage is conducted by the RAM module, which includes two blocks: a Region Construction Block (RCB) and a Region Interaction Block (RIB) first introduced in RANet. The inputs to this stage are the feature maps of F_s and F_t after a segmentation network forward pass. Figure 12 gives a detailed overview of the RCB and RIB building blocks.

Feature confidence domain adaptation module. Compared to the aforementioned domain adaptation model, our next module FCDAM mainly operates in the *feature confidence space*. After the model undergoes the alignment operation in feature- and output-space, FCDAM is used to further improve the confidence of domain-specific features given by the backbone architecture. Different from [214], the *entropy map* $E \in [0, 1]^{h \times w}$ is calculated by the given feature map. Thus, the loss of entropy map is:

$$\mathcal{L}_{ent}(F) = -\sum_{h,w} (\phi(F^{(h,w)}) \log(\phi(F^{(h,w)}))), \tag{6}$$

where ϕ is `Sigmoid` function applied at each pixel of feature map $F \in \mathbb{R}^{h \times w}$. During training G with the feature map F_s=$G(x_s)$ and F_t=$G(x_t)$ from source- and target

Methods	FS	OS	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
FANet	-	-	26.90	**62.98**	10.64	72.41	7.80	20.74	11.77	6.85	3.75	68.11	21.56	87.00	23.73	5.33	49.61	10.65	**0.54**	16.76	24.15	6.62
FANet	S	S	32.17	62.16	16.85	78.78	13.67	24.07	19.72	11.42	9.68	71.42	18.22	85.72	32.66	**11.75**	54.34	17.61	0.00	41.52	29.30	12.30
FANet	A	S	32.67	62.28	16.86	79.99	**17.64**	23.96	**19.78**	**12.33**	9.58	72.01	19.29	85.91	32.85	11.03	55.75	15.38	0.38	43.53	29.19	12.95
FANet	S+A	S	33.05	61.74	17.70	**80.07**	16.38	24.64	19.61	12.04	**9.79**	**72.27**	17.94	86.31	33.17	11.47	55.18	15.61	0.04	52.55	28.68	12.82
FANet	S+A	R	33.02	62.58	19.25	80.07	15.68	24.87	19.27	11.54	9.01	71.95	19.65	86.89	32.18	12.03	55.12	17.37	0.21	44.98	**29.93**	14.87
FANet	S+A+F	R	33.52	57.16	25.66	78.43	16.02	**26.88**	12.76	2.30	7.34	68.73	26.92	87.45	**36.51**	1.20	62.83	20.16	0.00	68.46	17.86	20.19
FANet-SSL	S+A	R	34.26	57.92	24.22	78.84	14.94	25.42	13.39	4.82	7.14	69.47	25.77	**87.92**	36.12	4.27	62.83	22.90	0.00	78.73	16.15	20.02
FANet-SSL	S+A+F	R	**35.67**	58.08	**28.75**	78.19	16.47	26.86	13.78	4.76	7.62	69.01	**34.58**	87.51	36.12	0.90	**64.06**	**27.50**	0.00	**84.99**	18.13	**20.35**

Table 1: Per-class results on DensePASS based on FANet [79] with different DA modules on our P2PDA framework. The size of input is 2048×400. S, A, R, and F represent SDAM, ADAM, RCDAM, and FCDAM respectively. Feature- and output-space are named as FS and OS for short. SSL represents the self-supervised learning with pseudo-labels. The first row is source-only.

Methods	FS	OS	Mean IoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
DANet	-	-	28.50	**70.68**	8.30	75.80	9.49	21.64	**15.91**	5.85	9.26	**71.08**	31.50	85.13	6.55	1.68	55.48	24.91	30.22	0.52	0.53	17.00
DANet	S	S	38.51	61.78	21.11	74.59	22.59	29.93	14.79	15.00	10.17	66.94	19.03	82.57	31.03	21.24	53.26	54.67	37.77	39.40	43.84	31.95
DANet	A	S	39.16	61.34	20.71	76.52	20.53	30.03	14.19	15.69	10.09	68.60	18.84	82.08	33.16	21.75	57.68	53.88	40.33	41.47	46.11	31.00
DANet	S+A	S	39.28	62.43	21.89	76.22	21.42	30.54	14.85	14.10	9.76	69.07	19.94	82.84	**34.56**	19.30	56.51	53.04	42.51	39.47	45.71	32.09
DANet	S	R	39.46	62.75	23.17	76.65	23.90	**30.82**	14.84	**18.44**	10.09	69.10	17.60	82.78	33.51	21.53	55.97	51.78	41.77	36.90	46.11	**32.12**
DANet	S+A	R	39.76	63.11	24.63	76.17	25.03	30.56	13.68	15.68	**10.53**	67.31	22.41	80.15	32.95	21.11	54.39	53.51	43.64	42.20	46.71	31.66
DANet	S+A+F	R	40.52	62.90	25.58	76.62	24.45	30.37	14.45	16.75	9.96	67.87	19.70	82.04	34.18	**22.95**	56.99	54.27	**44.15**	47.75	46.98	31.86
DANet-SSL	S+A	R	41.39	67.24	27.98	77.18	25.11	25.80	15.33	10.59	6.58	69.24	33.89	80.96	32.18	5.29	69.86	59.70	36.20	65.99	**47.47**	29.87
DANet-SSL	S+A+F	R	**41.99**	70.21	**30.24**	**78.44**	**26.72**	28.44	14.02	11.67	5.79	68.54	**38.20**	**85.97**	28.14	0.00	**70.36**	**60.49**	38.90	**77.80**	39.85	24.02
DANet*	S	R	41.35	68.38	**37.26**	75.51	26.28	31.81	15.62	8.99	10.33	66.22	31.74	80.68	33.69	16.81	64.81	47.67	28.05	61.81	44.92	34.98
DANet*	S+A	R	42.47	67.47	30.16	75.27	30.26	37.50	16.19	9.35	9.78	63.14	30.44	77.07	34.82	15.24	64.33	53.70	43.33	71.57	46.80	30.47
DANet*	S+A+F	R	42.87	66.92	29.97	77.34	**30.87**	**37.85**	15.04	11.12	9.60	62.80	31.03	78.08	36.27	18.01	63.66	54.83	42.86	74.22	45.96	28.13
DANet-SSL*	S+A	R	44.27	70.63	35.30	78.52	25.27	33.51	14.43	**13.80**	7.31	63.52	34.94	84.31	34.54	**19.08**	70.05	49.14	**48.80**	**75.11**	47.53	**35.36**
DANet-SSL*	S+A+F	R	**44.66**	**75.85**	34.21	**82.58**	28.75	35.58	**18.51**	12.65	**12.49**	**71.33**	**37.51**	**89.80**	**38.68**	15.99	**76.59**	**62.81**	12.25	61.56	**48.18**	33.26

Table 2: Per-class results on DensePASS based on DANet [61]. * means adding WildDash.

domain, FCDAM can improve the feature confidence by minimizing the loss of feature entropy maps. $\mathcal{L}_{adv}(G)$ for FCDAM in Eq. (5) is replaced by:

$$\mathcal{L}_{ent}(G(F)) = \lambda^{s}_{ent}\mathcal{L}_{ent}(G(F_s)) + \lambda^{t}_{ent}\mathcal{L}_{ent}(G(F_t)), \tag{7}$$

where both λ_{ent} are same as λ_{adv}. Note that, as shown in Figure 8, this FCDAM module eases the process of adding feature confidence learning to the original backbone without modification to the architecture of the whole domain adaptation framework.

3.1.4 *Experiments and Analysis*

3.1.4.1 *Ablation Studies for Segmentation Network*

As shown in Table 1, before adaptation, FANet [79] yields a mIoU of 26.90% indicating large room for improvement in cross-domain generalization. Our framework improves the result to 32.17% by using the SDAM module in both feature- and output-space (+5.27% gain). Integrating the attentional ADAM module also leads to a considerable boost (32.67% in mIoU, a +5.77% gain over the source-only baseline). A combination of our four modules yields the recognition result of 33.52% in mIoU. Furthermore, the pseudo-label self-supervised learning boosts our S+A+R and S+A+F+R adaptation results to 34.26% and 35.67% in mIoU, respectively.

Apart from using FANet, we conduct experiments with the accuracy-oriented segmentation network DANet [61] (in Table 2). The native source-trained DANet achieves a mIoU of only 28.50%, highlighting the sensitivity of modern segmentation networks to the Pinhole-to-Panoramic domain shift. The performance is strongly improved (+10.01% boost) through SDAM modules placed in feature- and output-space, achieving 38.51% in mIoU. Similarly, using the ADAM module yields a result of 39.16% (a +10.66% improvement over the source-only baseline). Combining both the SDAM and ADAM modules again slightly improves the performance (39.28% in mIoU). We further explore the use of the RCDAM module in output-space, yielding 39.46% mIoU (+10.96% boost over the baseline). The performance of 39.76% (+11.26% boost with respect to the original segmentation network) is achieved by combining three modules: SDAM, ADAM, and RCDAM. Integrating FCDAM leads 40.52% in mIoU (a +12.02% increase).

3.1.4.2 *Comparison with the State-of-the-Art*

Before delving into more comparisons, we introduce category definitions of ApolloScape [83], IDD [211], and Mapillary Vistas [142] datasets, in which the identical 19 categories following Cityscapes [44] can be obtained by class mapping. Models trained on ApolloScape [83] perform segmentation with 16 overlapping categories, where the *terrain*, *sky*, and *train* classes are discarded. The *train* class in IDD [211] and Mapillary Vistas [142] is excluded, thus other 18 classes are remained.

Next, we compare our approach with previous panoramic segmentation methods, including PASS [249] and ECANet [252]. Our P2PDA-driven DANet trained with Cityscapes and WildDash sources (detailed in Section 3.1.4.3) outperforms these works, achieving 44.66% by using pinhole data annotations only and successfully transferring beyond the FoV. Performing another run of the self-supervised learning stage elevates the mIoU to 48.52%, leading to the best segmentation result.

We now compare P2PDA with two state-of-the-art approaches for UDA: one method based on adversarial learning (CLAN [130]) and one built on self-training (CRST [311]), both adapting from Cityscapes to DensePASS. Our proposed framework clearly stands out in front of other domain adaptation pipelines, improving the performance by ~10%, showing the effectiveness of the attention-based design.

To broaden our comparison, we consider multi-supervision methods which benefit from multi-source data. Seamless-Scene-Segmentation [153] uses instance segmentation labels for auxiliary supervision, whereas USSS [94] performs multi-source semi-supervised learning. The outputs of these models are mapped to the 19 classes in DensePASS to be comparable with other models. ISSAFE [285] merges multiple training datasets including Cityscapes, KITTI-360 [114], and BDD [260] for safety-critical accident scene segmentation. Our experiments indicate that all these multi-source frameworks are sub-optimal in contrast to P2PDA which consistently leads to the best recognition rates. At the same time, P2PDA is trained on far less trainig data as the above approaches leverage larger databases, such as BDD/IDD and Mapillay, for training [252, 285]. Especially for the classes *building*, *truck*, *train*, and *bicycle*, our framework is a front-runner by a large margin, as seen in Table 3.

To grasp the key prediction differences before and after the domain adaptation with P2PDA, we compare the Pixel Accuracy (Acc) and IoU in different directions of the

Methods	Mean IoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
ERFNet [164]	16.65	63.59	18.22	47.01	9.45	12.79	17.00	8.12	6.41	34.24	10.15	18.43	4.96	2.31	46.03	3.19	0.59	0.00	8.30	5.55
PASS [249] (ERFNet)	23.66	67.84	28.75	59.69	19.96	29.41	8.26	4.54	8.07	64.96	13.75	33.50	12.87	3.17	48.26	2.17	0.82	0.29	23.76	19.46
ECANet (Omni-supervised) [252]	43.02	**81.60**	19.46	81.00	32.02	**39.47**	**25.54**	3.85	17.38	**79.01**	39.75	**94.60**	**46.39**	12.98	**81.96**	49.25	28.29	0.00	**55.36**	29.47
CLAN (Adversarial training) [130]	31.46	65.39	21.14	69.10	17.29	25.49	11.17	3.14	7.61	71.03	28.19	55.55	18.86	2.76	71.60	26.42	17.99	59.53	9.44	15.91
CRST-LRENT (Self-training) [311]	31.67	68.18	15.72	76.78	14.06	26.11	9.90	0.82	2.66	69.36	21.95	80.06	9.71	1.25	65.12	38.76	27.22	48.85	7.10	18.08
Seamless (Mapillary) [153]	34.14	59.26	24.48	77.35	12.82	30.91	12.63	**15.89**	**17.73**	75.61	33.30	87.30	19.69	4.59	63.94	25.81	**57.16**	0.00	11.59	19.04
USSS (IDD) [94]	26.98	68.85	5.41	67.39	15.10	21.79	13.18	0.12	7.73	70.27	8.84	85.53	22.05	1.71	58.69	16.41	12.01	0.00	23.58	13.90
SwiftNet (ApolloScape)	14.08	61.21	34.93	57.92	7.85	23.37	13.33	9.04	6.44	50.39	0.00	0.00	0.44	0.00	0.09	0.36	1.83	0.00	0.04	0.24
SwiftNet (Cityscapes) [146]	25.67	50.73	32.76	70.24	12.63	24.02	18.79	7.18	4.01	64.93	23.70	84.29	14.91	0.97	43.46	8.92	0.04	4.45	12.77	8.77
SwiftNet (KITTI-360)	25.00	69.03	27.71	68.07	15.70	16.26	15.29	0.00	4.43	64.71	31.01	84.86	23.02	0.00	45.08	9.72	0.00	0.00	0.00	0.00
SwiftNet (BDD)	24.69	4.26	25.11	74.16	15.53	22.74	11.70	0.00	10.58	70.86	26.55	92.26	25.12	0.00	58.78	31.35	0.00	0.00	0.00	0.00
SwiftNet (Merge3) [285]	32.04	68.31	**38.59**	81.48	15.65	23.91	20.74	5.95	0.00	70.64	25.09	90.93	32.66	0.00	66.91	42.30	5.97	0.07	6.85	12.66
Ours (Cityscapes)	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
Ours (Cityscapes+WildDash)	44.66	75.85	34.21	82.58	28.75	35.58	18.51	12.65	12.49	71.33	37.51	89.80	38.68	**15.99**	76.59	**62.81**	12.25	61.56	48.18	33.26
Ours* (Cityscapes+WildDash)	**48.52**	76.87	35.70	**85.16**	**33.93**	38.86	18.18	10.52	13.71	73.98	**41.89**	92.08	42.38	8.26	78.62	60.12	42.17	**81.21**	53.82	**34.49**

Table 3: Per-class results on DensePASS. Comparison with panoramic semantic segmentation, unsupervised domain adaptation, and multi-supervision methods. * denotes two rounds of SSL.

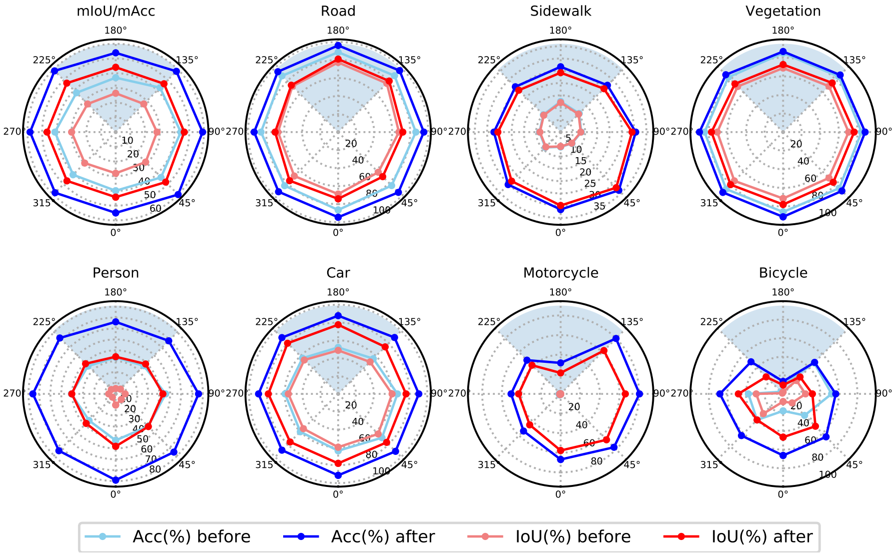


Figure 13: Class-wise Pixel Accuracy (Acc) and IoU comparison in different directions of the panoramic image, before and after adaptation.

panoramic image in Figure 13, where the blue-tinted regions indicate the section visible to a forward-facing narrow-FoV pinhole camera. We partition the 360° into 8 directions and compute the class-wise accuracy of navigation-critical categories separately for each direction. Our model leads to a considerable performance increase in all directions and for all the classes. While the same panoramic view can be achieved from multiple cameras surrounding a vehicle, our system enables reliable deployment using a single camera together with good performances in certain safety-critical directions. In particular, the recognition quality of *sidewalk*, *person*, and *motorcycle* is improved by an especially large margin through the domain adaptation paradigm. For the critical *road* and *car* segmentation relevant to autonomous driving, we have reached pixel accuracy at the level of 90% around the 360°.

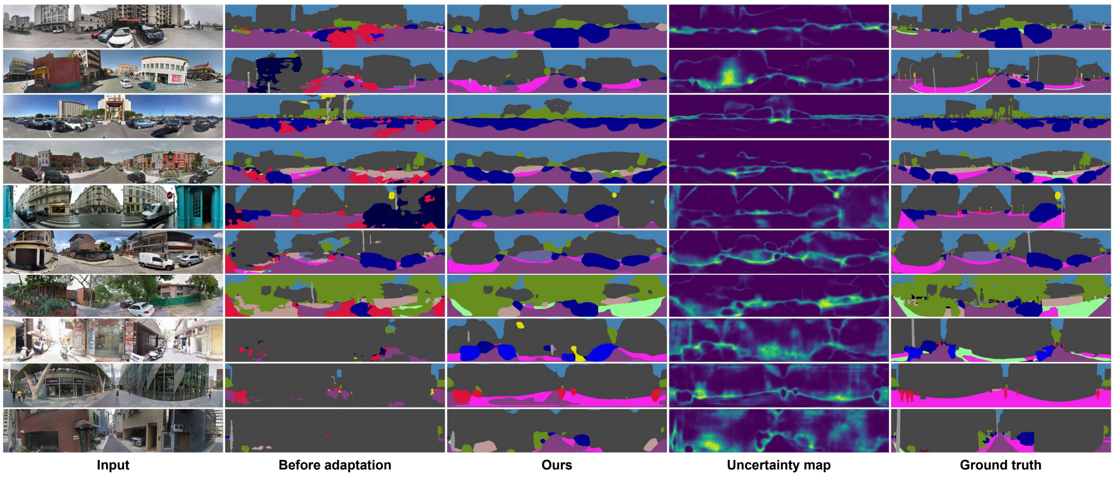


Figure 14: Qualitative examples of semantic segmentation on panoramic images. From left to right are input images, DANet predictions before adaptation, our results, uncertainty maps (brighter areas indicate higher uncertainty), and the ground truth. Zoom in for a better view.

3.1.4.3 *Comparison on Diverse Source Domains*

To further complement the image feature from these perspectives, we consider exploiting a more diverse dataset in the P2PDA framework. To achieve this, we leverage the WildDash dataset [266] with 4256 pinhole images, pixel-level annotations and more unstructured surroundings. For the training, we aggregate Cityscapes and WildDash sources without any complex joint training methodologies. As shown in the last rows of Table 2, we obtain better mIoU with the expanded training set, achieving 42.87% and 44.66% with different P2PDA variants.

3.1.4.4 *Qualitative Analysis*

We further demonstrate predictions of our adapted DANet in Figure 14, showing a clear performance decline of the source-only DANet when applied on panoramic images. In some of the top examples, the baseline often confuses the segmentation of some foreground categories, such as *cars*. Even in the last three lines, it cannot distinguish other categories from the *building* category in complex scenarios which is clearly better with the adapted DANet version. Despite lacking sharp boundaries, the adapted model is superior at distinguishing the categories, which is particularly important for autonomous vehicles. Strategies to augment the details include leveraging disentangled attention to handle detailed dependencies [252] or directly using detail-sensitive networks [97, 250] for adaptation. We want to mention that while the uncertainty maps are mainly used to select high-confident pseudo-labels for the self-supervised learning, they could also be utilized as an attention cue for the assistance system during driving.

3.2 DISTORTION-AWARE TRANSFORMERS FOR PANORAMAS

This section is based on our work published in *CVPR 2022* [283].

3.2.1 *Vision Transformer in Panoramas*

Panoramic semantic segmentation has received an increasing amount of attention in fields of Intelligent Transportation, facilitating a holistic and pixel-wise understanding of surrounding environments [134, 250, 280]. Due to the equirectangular projection [192, 252], panoramic camera data exhibit image distortions and object deformations (see Figure 15). Further, in the 360° image domain, labeled data is scarce which necessitates model training to be carried out on semantically matching narrow-FoV pinhole datasets. These two circumstances culminate in a significantly degraded performance on panoramic segmentation as compared to the pinhole counterpart [249]. Considering the intricacies of panoramas, convolution variants [43, 177, 203] and attention-augmented models [252] were proposed to mitigate image distortions and enlarge receptive fields of Convolutional Neural Networks (CNNs). However, they remain suboptimal in handling the severe deformations from pinhole- to panoramic data, and fail in establishing long-range contextual dependencies in the ultra-wide 360° images, which prove essential for accurate semantic segmentation [61, 296].

To address these challenges, we put forward a *Transformer for PAnoramic Semantic Segmentation (Trans4PASS)* architecture, and overcome image distortions and object deformations through two novel design choices: (1) Our Deformable Patch Embedding (DPE) is located at the early image sequentialization- and intermediate feature interpretation stages empowering the model to learn characteristic panoramic image distortions and preserve semantics. (2) Within the feature parsing stage, we introduce Deformable MLP (DMLP) module. This module is capable of mixing patches with learned spatial offsets, enhancing the ability for global context modeling.

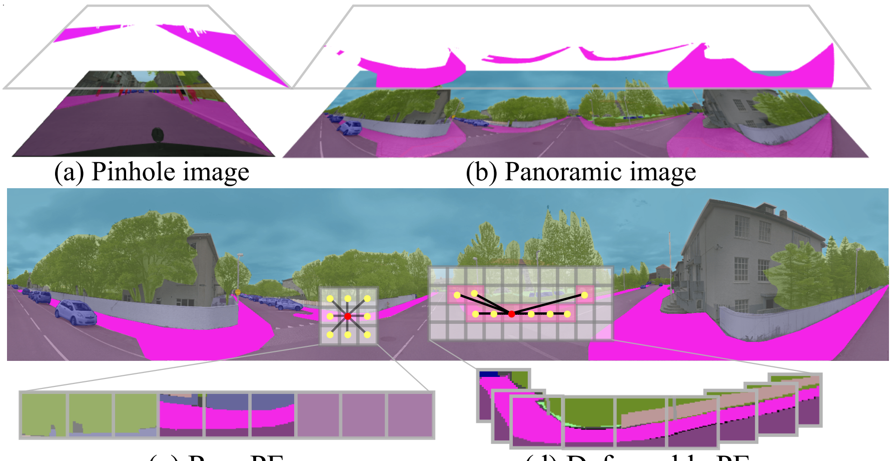
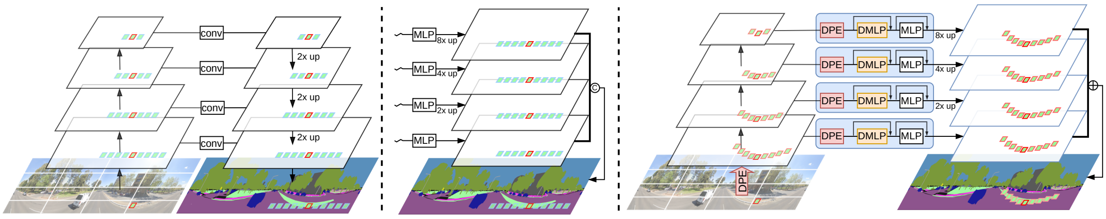

(a) Pinhole image
(b) Panoramic image
(c) Raw PE
(d) Deformable PE

Figure 15: Comparison of semantic segmentation with (a) narrow-angle pinhole image and (b) 360° panoramic image. Compared to (c) standard Patch Embeddings, our (d) Deformable Patch Embedding partitions 360° images while considering distortions, *e.g.* in *sidewalks*.



(a) FPN-like method (b) Vanilla-MLP method (c) Trans4PASS with DPE and DMLP

Figure 16: Comparison of segmentation transformers. Transformers (a) borrow a FPN-like decoder [296] from CNN counterparts or (b) adopt a vanilla-MLP decoder [238] for feature fusion, which lacks patch mixing. (c) *Trans4PASS* integrates Deformable Patch Embeddings (DPE) and Deformable MLP (DMLP) to handle distortions (see warped *terrain*) and mix patches.

3.2.2 *Trans4PASS Model*

To investigate the transformer model on panoramic semantic segmentation, we create two versions of *Trans4PASS* models (T: Tiny and S: Small). We build both with four stages, where for the tiny model, each stage encompasses 2 layers, for the small version the stages have {3, 4, 6, 3} layers. As shown in Figure 16, the pyramidal stages are inspired by recent transformers [222, 238], which reduce the feature scales in deeper layers. Given an input image with $H{\times}W{\times}3$, Trans4PASS makes use of a Patch Embedding (PE) module [238] to split the image into patches. To deal with the severe distortions in panoramas, a special *Deformable Patch Embedding (DPE)* module is proposed and applied in the encoder and decoder (Figure 16c). In the encoder, each feature map $\boldsymbol{f}_l{\in}\{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_4\}$ in the l^{th} stage is down-sampled by the l^{th} stride $\in\{4, 8, 16, 32\}$. The channel dimensions $C_l{\in}\{64, 128, 320, 512\}$ grow successively. Different from the FPN-like decoder [296] and vanilla-MLP based decoder [238] in Figure 16, we propose the *Deformable MLP (DMLP)* decoder structure, which mixes feature patches extracted via DPE. Given the extracted feature hierarchy in multiple scales from the encoder, four deformable decoder layers process the feature hierarchy into a consistent shape of $\frac{H}{4}{\times}\frac{W}{4}{\times}C_{emb}$, where we set the number of resulting embedding channels $C_{emb}{=}128$. An ensuing linear layer transforms the 128-channel output to contain the number of semantic classes of the respective task.

3.2.3 *Deformable Patch Embedding*

Spherical topological images captured by 360° cameras occupy a polar coordinate system with $\theta{\in}[0, 2\pi)$ and $\phi{\in}[0, \pi]$. To represent it in 2D space, the spherical data is usually converted into a panoramic format in euclidean-like space through the equirectangular projection. This process leads to severe shape distortions in the projected panoramic image, as seen in Figure 15. Therefore, a common PE module with fixed sampling positions does not respect these shape distortions of objects and the overall scene. Inspired by deformable convolution [47] and overlapping PE [238], we propose *Deformable Patch Embeddings (DPE)* and employ them on the input to the encoder and the decoder, splitting panoramic images and features. Given an input image or feature map $\boldsymbol{f}{\in}\mathbb{R}^{H{\times}W{\times}C_{in}}$, a standard PE module [52, 238] splits it into a flattened 2D patch sequence $\boldsymbol{z} \in \mathbb{R}^{(\frac{HW}{s^2}) \times (s^2 \cdot C_{in})}$, where $\frac{HW}{s^2}$ is the number of patches and s is the width and height of each patch. Each element in this sequence is passed through a linear projection layer transforming it into C_{out} dimensional embeddings.

Consider a single patch in $\boldsymbol{z}$ representing a rectangle of size $s \times s$ with s^2 positions. We can define a position offset relative to a location $(i, j) | i, j \in [1, s]$ in the patch as $\boldsymbol{\Delta}_{(i,j)} \in \mathbb{N}^2$. In standard PE, these offsets are fixed and lie in $\boldsymbol{\Delta}_{(i,j)} \in [\lfloor -\frac{s}{2} \rfloor, \lfloor +\frac{s}{2} \rfloor]^2$. Take *e.g.* a 3×3 patch, offsets $\boldsymbol{\Delta}_{(i,j)}$ relative to the center will lie in $[-1, 1] \times [-1, 1]$.

As we want to process panoramic images, which inherit distortions from the equirectangular projection, we can directly address this degradation in the PE. To this end, in our *Deformable Patch Embedding (DPE)*, we enable the model to learn a data-dependent offset $\boldsymbol{\Delta}^{DPE} \in \mathbb{N}^{H \times W \times 2}$ that can better cope with the spatial connections of objects, as present in distorted patches. DPE is learnable and predicts relative offsets based on the original input $\boldsymbol{f}$. The offset $\boldsymbol{\Delta}^{DPE}_{(i,j)}$ is calculated as depicted in Eq. (8).

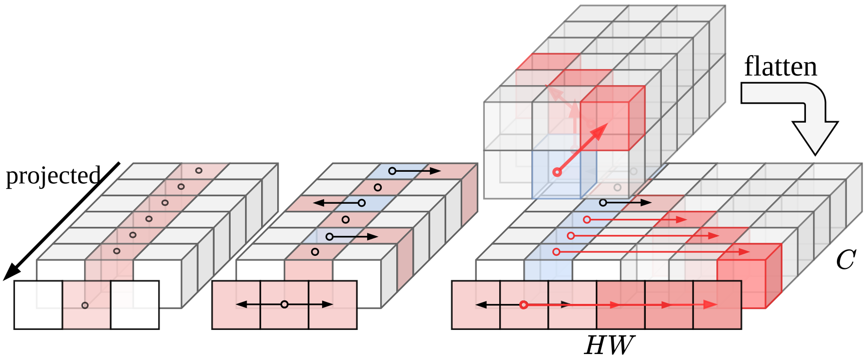
$$\boldsymbol{\Delta}^{DPE}_{(i,j)} = \begin{bmatrix} \min(\max(-\frac{H}{r}, g(\boldsymbol{f})_{(i,j)}), \frac{H}{r}) \\ \min(\max(-\frac{W}{r}, g(\boldsymbol{f})_{(i,j)}), \frac{W}{r}) \end{bmatrix}, \tag{8}$$

where $g(\cdot)$ is the offset prediction function, which we implement via the deformable convolution operation [47]. The hyperparameter r puts a constraint onto the offsets and is set as 4 in our experiments. The learned offsets make DPE adaptive and as a result distortion-aware.

3.2.4 *Deformable MLP*

Apart from the specific design of the encoder, the decoder with an adaptive feature parsing capacity is crucial in segmentation transformers [238, 281]. As shown in Figure 16a, some transformers [296] borrow a FPN-like decoder from the CNN counterpart [119], whose receptive field is limited to the feature resolution in its final stage [222]. SegFormer [238] takes inspiration from Multilayer Perceptron-based (MLP) models [206] and integrates a vanilla MLP to combine features (Figure 16b), but does not consider potential distortions in the imaging data. Next, we propose a mechanism to associate self-attention in Transformers and deformation-properties in 360° imagery. Linking both of these enables profiting from long-range dependencies for dense scene parsing and keeping this improvement when processing panoramic scenes. Achieving this distortion-aware property at manageable computational complexity, we put forward the *Deformable MLP (DMLP)* module. Within each stage of the decoder, DMLP mixes patches across the channel dimension, but with a particularly large receptive field, which improves the interpretation of features delivered by the aforementioned DPE.

Figure 17 shows the difference in MLP-based modeling: while the vanilla MLP (see Figure 17a) performs traditional linear projection without learning any spatial context, CycleMLP (see Figure 17b) has a limited spatial receptive field by hand-crafted, fixed offsets in mixing patches and their channels. In Figure 17c, the proposed DMLP generates a learned spatial offset (top) in a wider range and an adaptive manner. Given the input feature map $\boldsymbol{f} \in \mathbb{R}^{H \times W \times C_{in}}$, the spatial offset $\boldsymbol{\Delta}^{DMLP}_{(i,j,c)}$ is predicted channel-wise



(a) MLP (b) CycleMLP (c) DMLP

Figure 17: Comparison of MLP blocks. The spatial offsets of DMLP are learned adaptively from the input feature map.

as in Eq. (8) and is then flattened as $\boldsymbol{\Delta}^{\mathrm{DMLP}}_{(k,c)}$, where $k{\in}HW$ and $c{\in}C_{in}$, for mixing the flattened patch features $z{\in}\mathbb{R}^{HW\times C_{in}}$, as:

$$\hat{z}_{(k,c)} = \sum_{k=1}^{HW}\sum_{c=1}^{C_{in}} w^{T}_{(k,c)} \cdot z_{(k+\boldsymbol{\Delta}^{\mathrm{DMLP}}_{(k,c)},c)}, \tag{9}$$

where $w{\in}\mathbb{R}^{C_{in}\times C_{out}}$ is the weight matrix of a fully-connected (FC) layer. As shown in Figure 16c, the decoder has a similar structure as a MLP-Mixer block [206], consisting of DPE, DMLP, and MLP modules. The residual connections are kept. Formally, the four-stage decoder is denoted as:

$$\begin{aligned}
\hat{z}_l &= \mathbf{DPE}(C_l, C_{emb})(z_l), \forall_l{\in}\{1,2,3,4\}\\
\hat{z}_l &= \mathbf{DMLP}(C_{emb}, C_{emb})(\hat{z}_l) + \hat{z}_l, \forall_l\\
\hat{z}_l &= \mathbf{MLP}(C_{emb}, C_{emb})(\hat{z}_l) + \hat{z}_l, \forall_l\\
\hat{z}_l &= \mathbf{Up}(H/4, W/4)(\hat{z}_l), \forall_l\\
p &= \mathbf{LN}(C_{emb}, C_K)(\sum_{l=1} \hat{z}_l),
\end{aligned} \tag{10}$$

where $\mathbf{Up}(\cdot)$ and $\mathbf{LN}(\cdot)$ refer to the Upsample- and LayerNorm operations, and p is the prediction of K classes.

3.2.5 *Mutual Prototypical Adaptation*

We propose the *Mutual Prototypical Adaptation (MPA)* method to enable distilling knowledge via prototypes which we cultivate through source ground truth labels and target pseudo labels. Pseudo-labels depend on the few remaining mutual properties from pinhole and panoramic images, *e.g.*, scene distribution at the frontal viewing angle [42, 252]. Specifically, given the source (pinhole) domain with images and annotations $\mathcal{D}^s{=}\{(x^s, y^s)|x^s{\in}\mathbb{R}^{H\times W\times 3}, y^s{\in}\{0,1\}^{H\times W\times K}\}$ and the target (panoramic) domain $\mathcal{D}^t{=}\{(x^t)|x^t{\in}\mathbb{R}^{H\times W\times 3}\}$ without annotations, the goal of domain adaptation is to learn

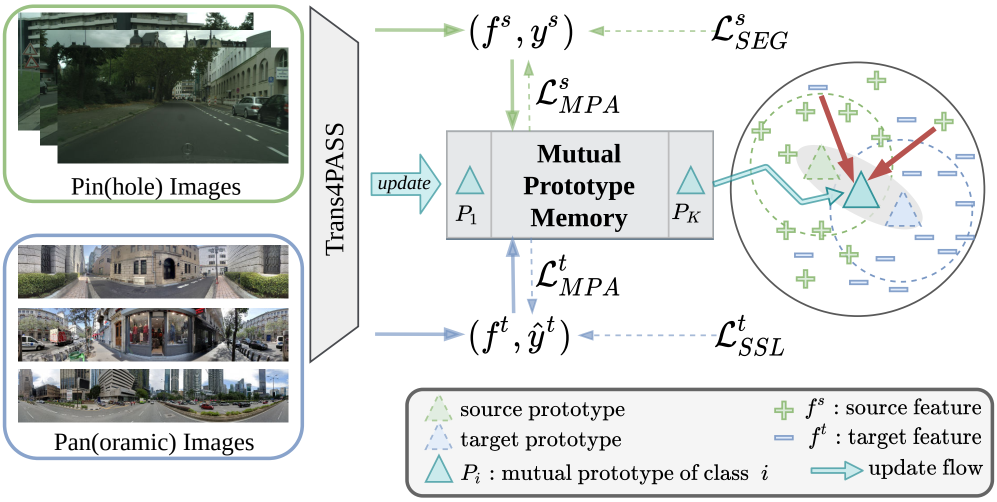


Figure 18: Diagram of mutual prototypical adaptation.

semantics from the source domain and transfer it to the target domain with K shared classes. The network is trained in $\mathcal{D}^s$ based on the segmentation loss:

$$\mathcal{L}^s_{SEG} = -\sum_{i,j,k=1}^{H,W,K} y^s_{(i,j,k)} \log(p^s_{(i,j,k)}), \tag{11}$$

where $p^s_{(i,j,k)}$ indicates the probability of pixel $x^s_{(i,j)}$ predicted as k-th class on the source domain. To generalize the source pre-trained model to the target data, a typical Self-Supervised Learning (SSL) scheme optimizes the model based on the pseudo labels $\hat{y}^t_{(i,j,k)}$ of pixels $x^t_{(i,j)}$ in the target domain:

$$\mathcal{L}^t_{SSL} = -\sum_{i,j,k=1}^{H,W,K} \hat{y}^t_{(i,j,k)} \log(p^t_{(i,j,k)}), \tag{12}$$

where the pseudo label is given by the most probable class in the model predictions: $\hat{y}^t_{(i,j,k)} = \mathbb{1}_{k \doteq \arg\max p^t_{(i,j,:)}}$. However, training with hard pseudo-labels leaves the model sensitive and fragile against errors in its own prediction and has only a limited positive effect on performance. Therefore, we advocate prototype-based alignment in the feature space, which brings two benefits: (1) it softens the hard pseudo-labels by using them in feature space instead of as direct targets and (2) it performs complementary alignment of semantic similarities in feature space.

Specifically, given a set with all n_s source feature maps and n_t target feature maps $\boldsymbol{F}{=}\{\boldsymbol{f}^s_1, \dots, \boldsymbol{f}^s_{n_s}\} \bigcup \{\boldsymbol{f}^t_1, \dots, \boldsymbol{f}^t_{n_t}\}$, with feature maps $\boldsymbol{f}$ fused from four-stage multi-scale features $\boldsymbol{f}{=}\sum_{l=1}^{4} f_l$. Each feature map is associated either with its respective source ground-truth label or a target pseudo-label. To compute the mutual prototype memory $\boldsymbol{\mathcal{M}}{=}\{P_1, ..., P_K\}$ with prototypes P_k we take the mean of all feature vectors (pixel-embeddings) from all feature maps in $\boldsymbol{F}$ that share the class label k. We initialize $\boldsymbol{\mathcal{M}}$ by computing the class-wise mean embeddings through the whole dataset and while training we update the prototype P_k at timestep t online by

$P_k^{t+1} \leftarrow m P_k^{t-1} + (1-m) P_k^t$ with a momentum m=0.999, where P_k^t is the mean pixel-embedding among embeddings that share the class-label k in the current mini-batch. An overview of this procedure is displayed in Figure 18. The mutual prototypical adaptation loss is inspired by the knowledge distillation loss [35], which drives the feature embedding $\boldsymbol{f}$ to be aligned with the prototypical feature map $\hat{\boldsymbol{f}}$ which is set up, by stacking the prototypes $P_k \in \boldsymbol{\mathcal{M}}$ according to the pixel-wise class distribution in either the source label or the pseudo-label. The resulting target $\hat{\boldsymbol{f}}$ has the same shape as $\boldsymbol{f}$. For brevity, only the source domain is displayed in Eq. (13), which is similar to the target domain.

$$\mathcal{L}_{MPA}^{s} = -\lambda \mathcal{T}^2 \mathbf{KL}(\phi(\hat{\boldsymbol{f}}^s/\mathcal{T}) \| \phi(\boldsymbol{f}^s/\mathcal{T})) - (1-\lambda)\mathbf{CE}(y^s, \phi(\boldsymbol{f}^s)), \tag{13}$$

where $\mathbf{KL}(\cdot)$, $\mathbf{CE}(\cdot)$, and $\phi(\cdot)$ are Kullback–Leibler divergence, Cross-Entropy, and Softmax function, respectively. The temperature $\mathcal{T}$ and hyper-parameter λ are 20 and 0.9 in our experiments.

The final loss is combined with a weight of α=0.001 as:

$$\mathcal{L} = \mathcal{L}_{SEG}^{s} + \mathcal{L}_{SSL}^{t} + \alpha(\mathcal{L}_{MPA}^{s} + \mathcal{L}_{MPA}^{t}). \tag{14}$$

3.2.6 *Experiments and Analysis*

3.2.6.1 *Pin2Pan Gaps*

Domain gap in outdoor scenarios. To quantify the Pin2Pan domain gap in outdoor scenarios, we evaluate over 15 off-the-shelf segmentation models trained on Cityscapes.[1] Table 4 summarizes the results tested on Cityscapes and DensePASS validation sets. Although previous transformers [238, 296] reduce the mIoU gap from ~50% of CNN-based counterparts to ~40%, the Pin2Pan gap remains large. The proposed Trans4PASS architecture has a high performance on pinhole image segmentation and also outperforms other methods on panoramic segmentation with 44.8% mIoU without any adaptation strategy. It indicates that distortion-aware features and long-range cues maintained in both low and high levels of Transformers as opposed to the context learned in higher-levels of CNNs, are important for wide-FoV panoramic segmentation.
Domain gap in indoor scenarios. Table 5 shows Pin2Pan domain gaps in indoor scenarios. As pinhole and panoramic images from Stanford2D3D are captured under the same setting, the Pin2Pan gap is smaller compared to the outdoor scenario. Still, in light of other CNN- and transformer-based methods, the small Trans4PASS version achieves 50.20% and 48.34% mIoU in pinhole- and panoramic image segmentation, yielding the smallest performance drop.

3.2.6.2 *Trans4PASS Structural Analysis*

Effect of DPE. We compare DPE against DePatch from DPT [37]. While the object-aware offsets and scales in DPT make patches shift around the object, our DPE is flexible to split image patches and is decoupled from object proposals. As shown in the

[1] MMSegmentation: https://github.com/open-mmlab/mmsegmentation.

Network	Backbone	CS	DP	Gaps
SwiftNet [146]	ResNet-18	75.4	25.7	-49.7
Fast-SCNN [155]	Fast-SCNN	69.1	24.6	-44.5
ERFNet [164]	ERFNet	72.1	16.7	-55.4
FANet [79]	ResNet-34	71.3	26.9	-44.4
PSPNet [293]	ResNet-50	78.6	29.5	-49.1
OCRNet [262]	HRNetV2p-W18	78.6	30.8	-47.8
DeepLabV3+ [29]	ResNet-101	80.9	32.5	-48.4
DANet [61]	ResNet-101	80.4	28.5	-51.9
DNL [256]	ResNet-101	80.4	32.1	-48.3
Semantic-FPN [97]	ResNet-101	75.8	28.8	-47.0
ResNeSt [275]	ResNeSt-101	79.6	28.8	-50.8
OCRNet [262]	HRNetV2p-W48	80.7	32.8	-47.9
SETR-Naive [296]	Transformer-L	77.9	36.1	-41.8
SETR-MLA [296]	Transformer-L	77.2	35.6	-41.6
SETR-PUP [296]	Transformer-L	79.3	35.7	-43.6
SegFormer-B1 [238]	SegFormer-B1	78.5	38.5	-40.0
SegFormer-B2 [238]	SegFormer-B2	81.0	42.4	-38.6
Trans4PASS-T	Trans4PASS-T	79.1	41.5	-37.6
Trans4PASS-S	Trans4PASS-S	81.1	**44.8**	-36.3

Table 4: Performance gaps of CNN- and transformer-based models from Cityscapes (**CS**) @ 1024×512 to DensePASS (**DP**).

Network	Backbone	SPin	SPan	Gaps
Fast-SCNN [155]	Fast-SCNN	41.71	26.86	-14.85
SwiftNet [146]	ResNet-18	42.28	34.95	-7.87
DANet [61]	ResNet-50	43.33	37.76	-5.57
DANet [61]	ResNet-101	40.09	31.81	-8.28
Trans4Trans-T [281]	PVT-T	41.28	24.45	-16.83
Trans4Trans-S [281]	PVT-S	44.47	23.11	-21.36
Trans4PASS-T	Trans4PASS-T	49.05	46.08	-2.97
Trans4PASS-S	Trans4PASS-S	50.20	**48.34**	-1.86

Table 5: Performance gaps from Stanford2D3D-Pinhole (**SPin**) dataset to Stanford2D3D-Panoramic (**SPan**) dataset on fold-1.

Network	Encoder	Decoder	GFLOPs	#P	CS	DP
(1) Compare PEs and MLPs:						
Trans4PASS	MiT-B1*	DMLP	13.11	13.10	69.48	36.50
Trans4PASS	MiT-B1†	CycleMLP	9.83	13.60	73.49	40.16
Trans4PASS	MiT-B1†	ASMLP	13.40	14.19	73.65	42.05
Trans4PASS	MiT-B1†	DMLP	12.02	13.93	72.49	45.89 (+9.39)
(2) Compare encoders and decoders:						
PVT [222]	PVT-T	FPN	11.17	12.76	71.46	31.20
PVT [222]	PVT-T	Vanilla MLP	14.56	12.84	70.60	32.85
PVT [222]	PVT-T	DMLP	13.11	13.10	71.75	35.18 (+3.98)
Trans4PASS	PVT-T†	DMLP	13.18	13.10	69.62	36.50 (+5.30)
SegFormer [238]	MiT-B1	Vanilla MLP	13.27	13.66	74.93	39.02
SegFormer [238]	MiT-B1	FPN	9.88	13.58	73.96	41.14
SegFormer [238]	MiT-B1	DMLP	11.82	13.92	73.10	45.14 (+6.12)
Trans4PASS	MiT-B1†	DMLP	12.02	13.93	72.49	**45.89** (+6.87)

Table 6: Trans4PASS structural analysis. * and † denote DPT [37] and our DPE. **#P**: #Parameters in millions. Models are from Cityscapes (**CS**) @ 512×512 to DensePASS (**DP**) @ 2048×400.

first group of Table 6, compared with DPT, our DPE-based Trans4PASS adds +3.01% and +9.39% mIoU on Cityscapes and DensePASS, respectively.

Effect of DMLP. To ablate the effect of different MLP-like modules embedded in the decoder of Trans4PASS, we substitute DMLP by CycleMLP [34] and ASMLP [112] modules. DMLP is lighter than ASMLP with fewer GFLOPs, parameters and it is more adaptive as opposed to the fixed offsets in CycleMLP. The first group of Table 6 shows that DMLP outperforms both modules with 3% to 5% in mIoU.

Effect of encoders and decoders. With the same encoder as PVT, a DMLP-based decoder brings a +3.98% improvement compared to the FPN- and MLP-based decoders, as shown in the second group of Table 6. When our DPE is applied in the early stage of the PVT encoder, further improvements of +5.30% can be made. Similar improvement results (+6.12% and +6.87%) are evident in experiments with a SegFormer encoder. Overall, these results show that DPE and DMLP can be integrated into diverse backbones, significantly improving distortion-adaptability for panoramic segmentation.

3.2.6.3 *Pin2Pan Adaptation*

Ablations in outdoor scenarios. To verify the generalization ability of applying Trans4PASS in adaptation methods, FANet and DANet used in P2PDA [134] are replaced by Trans4PASS-T/-S, as visible in Table 7b. Trans4PASS brings >10% performance gains due to the captured long-range contexts and distortion-aware features. Without the advantage of a superior network architecture, MPA achieves 51.93% and 54.77% with Trans4PASS-T and -S models, surpassing 51.05% and 52.91% of P2PDA. The second and third ablation groups of Table 7b show how Trans4PASS-T and -S match up against each other. Individually, MPA is on par with the SSL-based method. When combining both, MPA and SSL, Trans4Pass-S obtains new state-of-the-art performance

Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
ERFNet [164]	16.65	63.59	18.22	47.01	9.45	12.79	17.00	8.12	6.41	34.24	10.15	18.43	4.96	2.31	46.03	3.19	0.59	0.00	8.30	5.55
PASS (ERFNet) [249]	23.66	67.84	28.75	59.69	19.96	29.41	8.26	4.54	8.07	64.96	13.75	33.50	12.87	3.17	48.26	2.17	0.82	0.29	23.76	19.46
ECANet (Omni-supervised) [252]	43.02	**81.60**	19.46	81.00	**32.02**	39.47	25.54	3.85	17.38	79.01	39.75	**94.60**	46.39	12.98	**81.96**	49.25	28.29	0.00	55.36	29.47
CLAN (Adversarial) [130]	31.46	65.39	21.14	69.10	17.29	25.49	11.17	3.14	7.61	71.03	28.19	55.55	18.86	2.76	71.60	26.42	17.99	59.53	9.44	15.91
CRST (Self-training) [311]	31.67	68.18	15.72	76.78	14.06	26.11	9.90	0.82	2.66	69.36	21.95	80.06	9.71	1.25	65.12	38.76	27.22	48.85	7.10	18.08
P2PDA (Adversarial) [280]	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
SIM (Self-training) [227]	44.58	68.16	32.59	80.58	25.68	31.38	23.60	19.39	14.09	72.65	26.41	87.88	41.74	16.09	73.56	47.08	42.81	56.35	47.72	39.30
PCS (Self-training) [264]	53.83	78.10	**46.24**	86.24	30.33	**45.78**	34.04	22.74	13.00	**79.98**	33.07	93.44	47.69	22.53	79.20	61.59	67.09	83.26	58.68	39.80
USSS (IDD) [94]	26.98	68.85	5.41	67.39	15.10	21.79	13.18	0.12	7.73	70.27	8.84	85.53	22.05	1.71	58.69	16.41	12.01	0.00	23.58	13.90
USSS (Mapillary) [94]	30.87	71.01	31.85	76.79	12.13	23.61	11.93	3.23	10.15	73.11	31.24	89.59	16.05	3.86	65.27	24.46	18.72	0.00	9.08	14.48
Seamless (Mapillary) [153]	34.14	59.26	24.48	77.35	12.82	30.91	12.63	15.89	17.73	75.61	33.30	87.30	19.69	4.59	63.94	25.81	57.16	0.00	11.59	19.04
SwiftNet (Cityscapes) [146]	25.67	50.73	32.76	70.24	12.63	24.02	18.79	7.18	4.01	64.93	23.70	84.29	14.91	0.97	43.46	8.92	0.04	4.45	12.77	8.77
SwiftNet (Merge3) [285]	32.04	68.31	38.59	81.48	15.65	23.91	20.74	5.95	0.00	70.64	25.09	90.93	32.66	0.00	66.91	42.30	5.97	0.07	6.85	12.66
Trans4PASS-S (ours)	**55.25**	78.39	41.62	**86.47**	31.56	45.47	34.02	22.98	18.33	79.63	41.35	93.80	49.02	22.99	81.05	67.43	**69.64**	86.04	60.85	39.20
Trans4PASS-S (ours)*	**56.38**	79.91	42.68	86.26	30.68	42.32	**36.61**	**24.81**	**19.64**	78.80	**44.73**	93.84	**50.71**	**24.39**	81.72	**68.86**	66.18	**88.62**	**63.87**	**46.62**

(a) Per-class results on DensePASS. Comparison with state-of-the-art panoramic segmentation, domain adaptation, and multi-supervision methods. * means multi-scale (MS) evaluation.

Network	Method	mIoU(%)
FANet	P2PDA	35.67
DANet	P2PDA	41.99
Trans4PASS-T	P2PDA	51.05
Trans4PASS-S	P2PDA	52.91
Trans4PASS-T	-	45.89
Trans4PASS-T	Warm-up	50.56
Trans4PASS-T	SSL	51.86
Trans4PASS-T	MPA	51.93
Trans4PASS-T	MPA + SSL	53.26
Trans4PASS-T	MPA + SSL + MS	**54.72**
Trans4PASS-S	-	48.73
Trans4PASS-S	Warm-up	52.59
Trans4PASS-S	SSL	54.67
Trans4PASS-S	MPA	54.77
Trans4PASS-S	MPA + SSL	55.25
Trans4PASS-S	MPA + SSL + MS	**56.38**

(b) Results on DensePASS.

Network	Method	mIoU(%)
DANet	-	40.28
DANet	P2PDA	42.26
PVT-Tiny	-	24.45
PVT-Tiny	P2PDA	39.66
PVT-Small	-	23.11
PVT-Small	P2PDA	43.10
Trans4PASS-T	-	46.08
Trans4PASS-T	MPA	47.48
Trans4PASS-S	-	48.34
Trans4PASS-S	MPA	**52.15**
DANet	Supervised	44.15
Trans4PASS-S	Supervised	**53.31**

(c) Results on SPan @ fold-1.

	Method	Input	mIoU(%)
Supervised	StdConv [203]	RGB	32.6
	CubeMap [203]	RGB	33.8
	DistConv [203]	RGB	34.6
	UNet [166]	RGB-D	35.9
	GaugeNet [43]	RGB-D	39.4
	UGSCNN [86]	RGB-D	38.3
	HexRUNet [269]	RGB-D	43.3
	Tangent [58] (ResNet-101)	RGB	45.6
	HoHoNet [192] (ResNet-101)	RGB	52.0
	Trans4PASS (Small)	RGB	52.1
	Trans4PASS (Small+MS)	RGB	**53.0**
UDA	Trans4PASS (Source only)	RGB	48.1
	Trans4PASS (MPA)	RGB	50.8
	Trans4PASS (MPA+MS)	RGB	**51.2**

(d) Results on SPan @ 3 folds.

Table 7: Results and studies of Pin2Pan domain adaptation in indoor and outdoor scenarios.

on DensePASS, reaching 55.25% in mIoU and 56.38% with multi-scale evaluation. This verifies that MPA works collaboratively with pseudo labels and provides a complementary feature alignment incentive.

Omnidirectional segmentation. To showcase the effectiveness of MPA on omnidirectional segmentation, the panoramic image is divided into 8 directions and evaluated individually. The polar diagram in Figure 19 demonstrates that MPA brings uniform improvement to omnidirectional segmentation. Apart from benefiting the stuff classes (*road*, *sidewalk*, and *terrain*), MPA improves the segmentation of object classes, such as *person* and *truck*. Due to the panorama boundary at 180°, IoUs of *motorcycle* and *bicycle* are impacted, still consistent and large accuracy boosts with MPA in all directions for different classes are observed.

Comparison with outdoor methods. In Table 7a, we compare our solution with recent panoramic segmentation [249, 252] and domain adaptation [130, 227, 264, 280, 311] methods. Following [280], we also involve multi-supervision methods [94, 153, 285] which require much more data, to broaden the comparison. MPA-Trans4PASS arrives at the highest mIoU of 56.38%, outperforming the previous best P2PDA-SSL on DensePASS by 14.39% and the prototypical method [264] adapted by Trans4PASS. Trans4PASS obtains top scores on 10 of 19 classes. Notably, our solution shows improvements on challenging categories, *e.g.*, *truck*, *train*, *motorcycle*, and *bicycle*.

Adaptation results in indoor scenarios. The experiments in Table 7c are conducted according to the fold-1 data splitting [5] on the Stanford-Panoramic dataset. Our MPA surpasses the previous state-of-the-art P2PDA with DANet and it is even better than

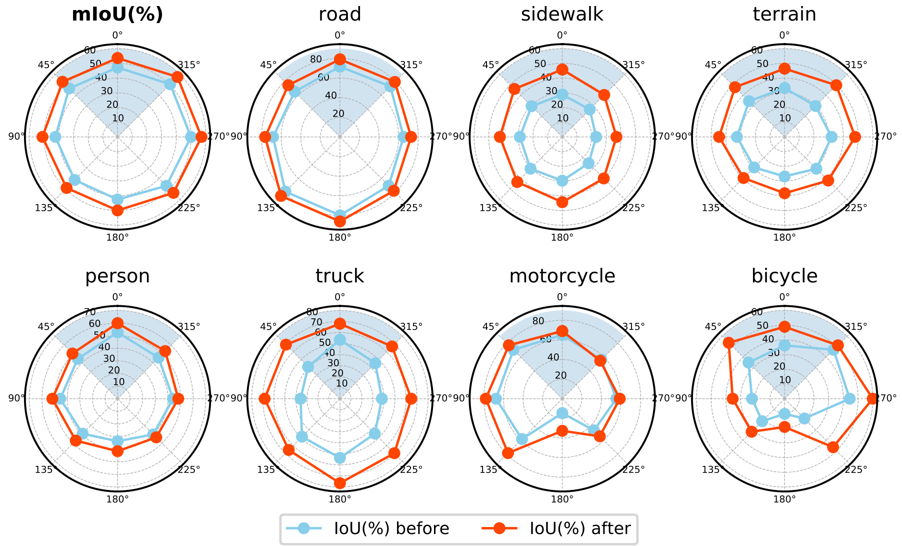


Figure 19: Comparison of omnidirectional segmentation before and after our MPA.

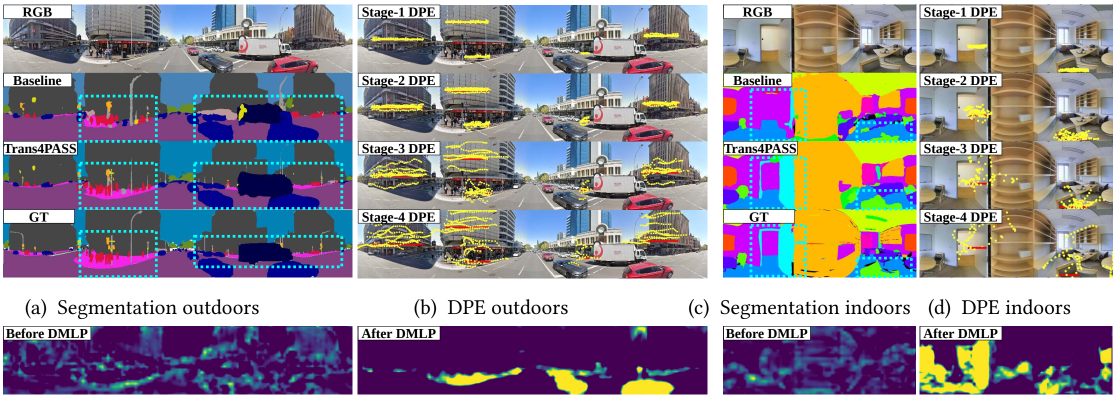
the one adapted by a PVT-Small backbone. Overall, our Trans4PASS-S with MPA achieves the highest mIoU (52.15%), even reaching the level of the fully-supervised Trans4PASS-S (53.31%) which does have access to panoramic image annotations.

Comparison with indoor methods. Before and after adaptation in Table 7d, our Trans4PASS-S model (~14M parameters) obtains a high mIoU score (51.2%), even comparable to existing fully-supervised and transfer-learning methods, which are based on ResNet-101 backbones (~44M parameters and 52.0% mIoU).

3.2.6.4 *Qualitative Analysis*

Panoramic segmentation visualizations. Figure 20a and Figure 20c demonstrate that Trans4PASS handles the distortion of panoramic images very well as compared to indoor [222] and outdoor [238] baseline models. Especially, the segmentation results for *sidewalks* and *pedestrians* from Trans4PASS have more accurate classifications and boundary distinctions, while the baseline model is confused by the distorted shape and space, due to the lacking capacity to learn long-range contexts and distortion-aware features. In the indoor case of Figure 20c, the *door* and *chair* categories are barely detected by the baseline model, but our Trans4PASS can output precise segmentation masks on both objects.

DPE and DMLP visualizations. Figure 20b and Figure 20d visualize effects of Deformable PE from four stages of Trans4PASS. The red dots denote the centers of a selected patch (size of $s \times s$) sequence. Given learned offsets from DPE, s^2 yellow sampling dots are shifted to semantic-relevant areas in a flexible way, where each pixel is adaptive to distorted objects and space, like the deformed *building* and *sidewalk* (see Stage-4 DPE in Figure 20b). Besides, to verify the effect of Deformable MLP, two feature map pairs from the 75^{th} channel before and after DMLP are displayed in Figure 20e and 20f. The feature maps (indoors/outdoors) after DMLP present semantically recognizable responses, *e.g.* on regions of distorted *sidewalks* or *doors*, as compared to those before the DMLP module.



(a) Segmentation outdoors (b) DPE outdoors (c) Segmentation indoors (d) DPE indoors

(e) DMLP outdoors (f) DMLP indoors

Figure 20: Qualitative comparisons, DPE and DMLP visualizations. (a) and (c) are predictions, where the baseline has neither DPE/DMLP nor MPA. The ● dots in (b) and (d) are sampling points shifted by learned offsets *w.r.t.* the ● patch center of DPE (from decoder). (e) and (f) show the #75 channel maps of stage-3 before and after DMLP. Zoom in for better view.

3.3 CHAPTER CONCLUSION

Omnidirectional scene understanding through panoramic semantic segmentation is a challenging task due to the large field-of-view and the presence of distortions in panoramic images. In this chapter, we propose two novel techniques for panoramic semantic segmentation, as the first research theme in the field of ITS:

- **Pinhole-to-Panoramic Domain Adaptation (P2PDA)**: A new setting of unsupervised domain adaptation (UDA) is investigated, with the aim of transferring model from the label-rich pinhole image domain to the label-scare panoramic image domain, and benchmarking panoramic semantic segmentation.
- **Distortion-aware Transformers for Panoramas (Trans4PASS)**: This research mainly investigates the potential of the long-range modeling ability of vision transformers to address the challenges of high distortion and deformation in panoramic images, and study a unified solution to address indoor and outdoor scenes.

Here is a more detailed overview of the contributions of each section in this chapter:

Contribution 1: We construct a new dataset, *DensePASS*, for benchmarking panoramic semantic segmentation. The dataset provides unlabelled panoramas for domain adaptation, along with a test set containing pixel-wise manual labels.

Contribution 2: We propose a novel unsupervised domain adaptation framework, *P2PDA*, to transfer models from pinhole to panoramic image domain by effectively addressing the domain shift between pinhole and panoramic images.

Contribution 3: We create a novel distortion-aware vision transformer for panoramic semantic segmentation, *Trans4PASS*, to handle the distortions in panoramic images and achieve state-of-the-art results on both the indoor and outdoor 360° datasets.

4

TOWARDS ROBUST SCENE UNDERSTANDING

In this chapter, we present the second research theme in ITS, *i.e.*, *robust scene understanding* through multimodal semantic segmentation. The challenge lies in establishing a unified fusion mechanism for diverse sensory data, such as commonly-used RGB, Depth, LiDAR, Event, Thermal, Polarization, and more. To achieve this, our contributions are two-fold: (1) We pioneer a unified cross-modal fusion model for RGB-X semantic segmentation, *i.e.*, *CMX*. It is presented in Section 4.1, based on our work published in *Transactions on ITS 2023* [278]. (2) To further enhance robustness, an advanced version, *i.e.*, *CMNeXt*, is designed for arbitrary-modal semantic segmentation, capable of fusing up to 80 modalities. Besides, we create a novel dataset DeLiVER that involves 4 modalities, 5 weather conditions, and 4 different sensor failure cases. This dataset is presented in Section 4.2, based on our *CVPR 2023* publication [279].

4.1 CROSS-MODAL FUSION FOR RGB-X SEMANTIC SEGMENTATION

This section is based on our work published in *Transactions on ITS 2023* [278].

4.1.1 *RGB-X Fusion Paradigms*

Robust scene understanding is essential for safe autonomous driving in Intelligent Transportation Systems (ITS) [285]. Thanks to the development of sensor technologies, there is a growing variety of modular sensors which are highly applicable for ITS applications. Different types of sensors can supply RGB images with rich complementary information (see Figure 21). For example, *depth* measurement can help identify the boundaries of objects and offer geometric information of dense scene elements [36, 82]. *Thermal* images facilitate to discern different objects through their specific infrared imaging [71, 287]. Besides, *polarimetric*- and *event* information are advantageous for perception in specular- and dynamic real-world scenes [235, 285]. *LiDAR* data can provide spatial information in driving scenarios [309]. Thereby, a research question arises: *How to construct a unified model to incorporate the fusion of RGB with various modalities,* i.e., *RGB-X semantic segmentation as illustrated in Figure 21?*

Existing multimodal semantic segmentation methods can be divided into two categories: (1) The first category [20, 31] employs a single network to extract features from RGB and another modality, which are fused in the input stage (see Figure 22a). (2) The second type of approaches [36, 48, 287] deploys two backbones to perform feature

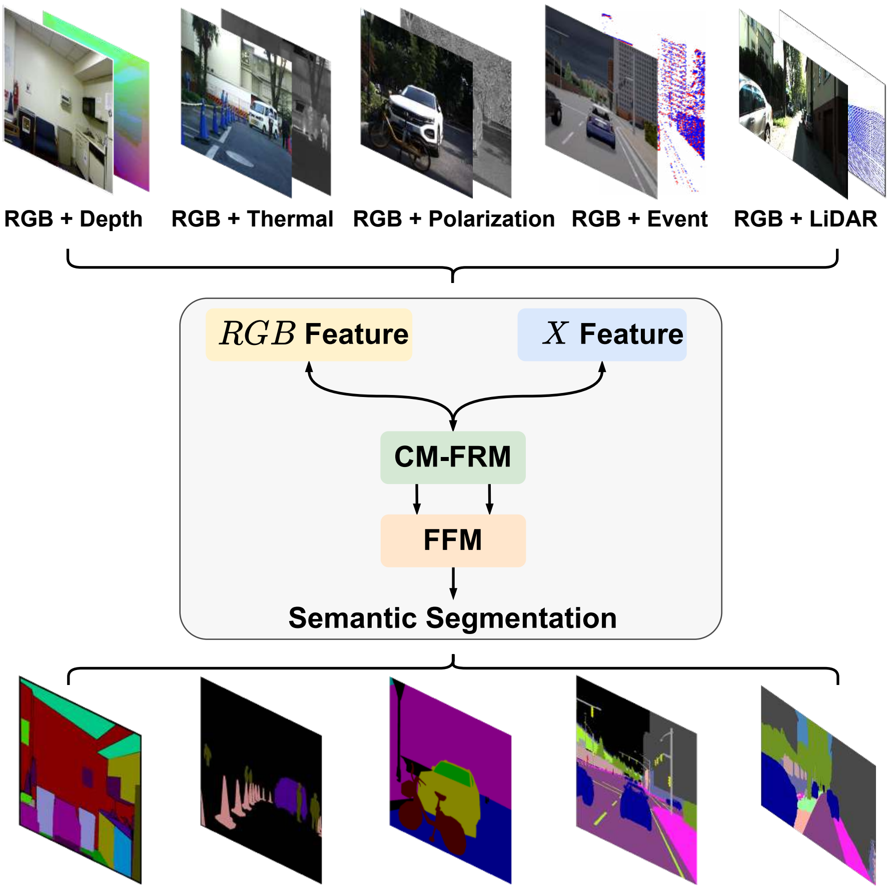
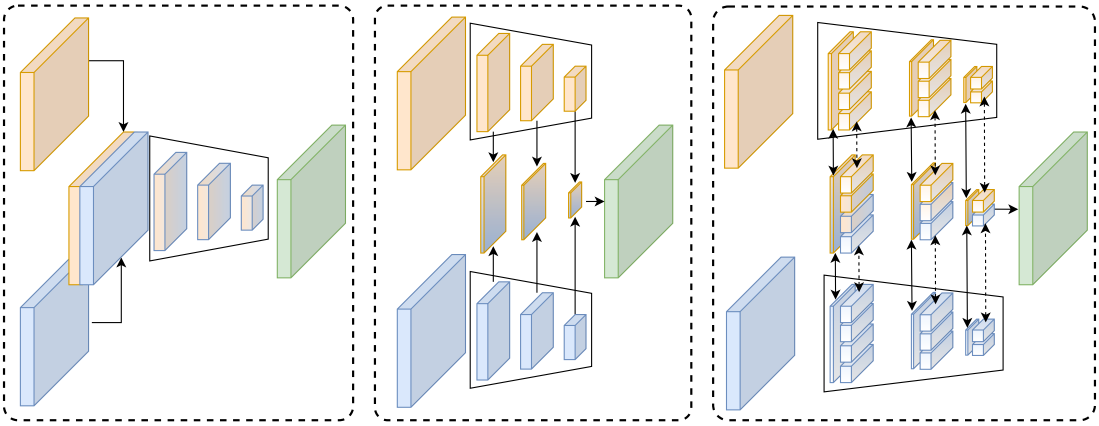


Figure 21: RGB-X semantic segmentation unifies diverse sensing modality combinations: RGB-Depth, -Thermal, -Polarization, -Event, and -LiDAR segmentation. CMX is established with Cross-Modal Feature Rectification Module (*CM-FRM*) to calibrate the features of RGB- and X-modality and Feature Fusion Module (*FFM*) to perform the exchange of long-range context and combine features for RGB-X semantic segmentation.

extraction from RGB- and another modality separately then fuses the extracted two features into one feature for semantic prediction (see Figure 22b). However, both types are usually well-tailored for a single specific modality pair (*e.g.*, RGB-D or RGB-T), yet hard to be extended to operate with other modality combinations. For example, regarding our observation in Figure 23, ACNet [82] and SA-Gate [36], designed for RGB-D data, perform less satisfactorily in RGB-T tasks. To flexibly cover various sensor combinations for ITS applications, a unified *RGB-X semantic segmentation*, is desirable and advantageous. Its benefits are two-fold: (1) It can save research and engineering efforts, with no need to adapt architectures for a specific modality combination scenario. (2) It enables that a system equipped with multimodal sensors can readily leverage new sensors when they become available [66, 193], which is conducive to robust scene perception. For this purpose, in this work, we spend efforts to construct a modality-agnostic framework for unified RGB-X semantic segmentation.

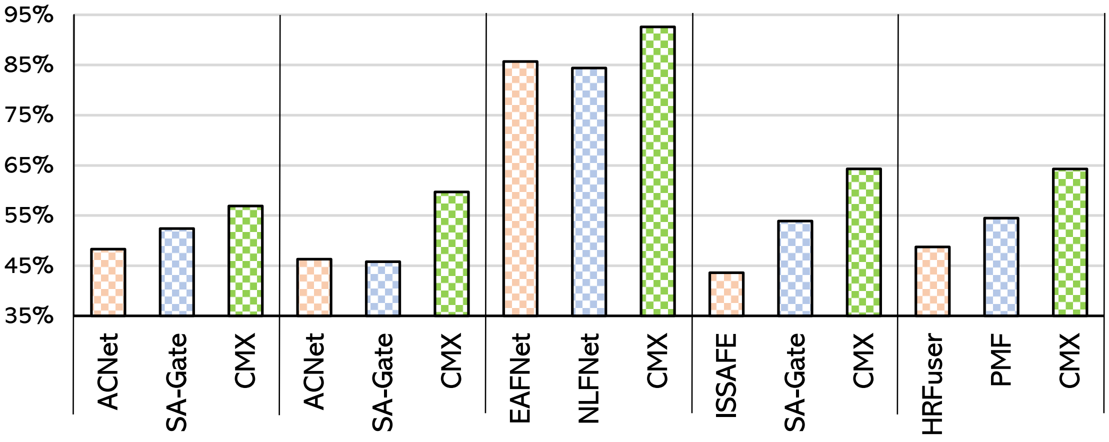
Compared to existing multimodal fusion modules [48, 82, 235] based on ConvNets, it remains unclear whether potential improvements on RGB-X semantic segmentation can be materialized via vision transformers [52, 125, 207, 212]. Crucially, while some



(a) Input fusion (b) Feature fusion (c) Interactive fusion

Figure 22: Comparison of different fusion methods. (a) Input fusion merges inputs with modality-specific operations [20, 31]. (b) Feature fusion applies channel attention to fuse features in a unidirectional manner [36, 82]. (c) Our interactive fusion incorporates bidirectional cross-modal feature rectification, and sequence-to-sequence cross-attention, yielding comprehensive cross-modal interactions.

previous works [36, 82] use a simple global multimodal interaction strategy, it does not generalize well across different sensing data combinations [287]. We hypothesize that for RGB-X semantic segmentation with various supplements and uncertainties, comprehensive cross-modal interactions should be provided, to fully exploit the potential of cross-modal complementary features. To tackle the aforementioned challenges, we propose *CMX*, a universal cross-modal fusion framework for RGB-X semantic segmentation in an interactive fusion manner (Figure 22c). Specifically, CMX is built as a two-stream architecture, *i.e.*, RGB- and X-modal streams.



(a) RGB-D (b) RGB-T (c) RGB-P (d) RGB-E (e) RGB-L

Figure 23: Performance comparison on different RGB-X semantic segmentation benchmarks. SA-Gate [36] designed for RGB-D data (*e.g.*, on NYU Depth V2 dataset [185]), is less effective on RGB-T or RGB-E tasks. Our modality-agnostic CMX, for the first time, outperforms modality-specific methods on five segmentation tasks.

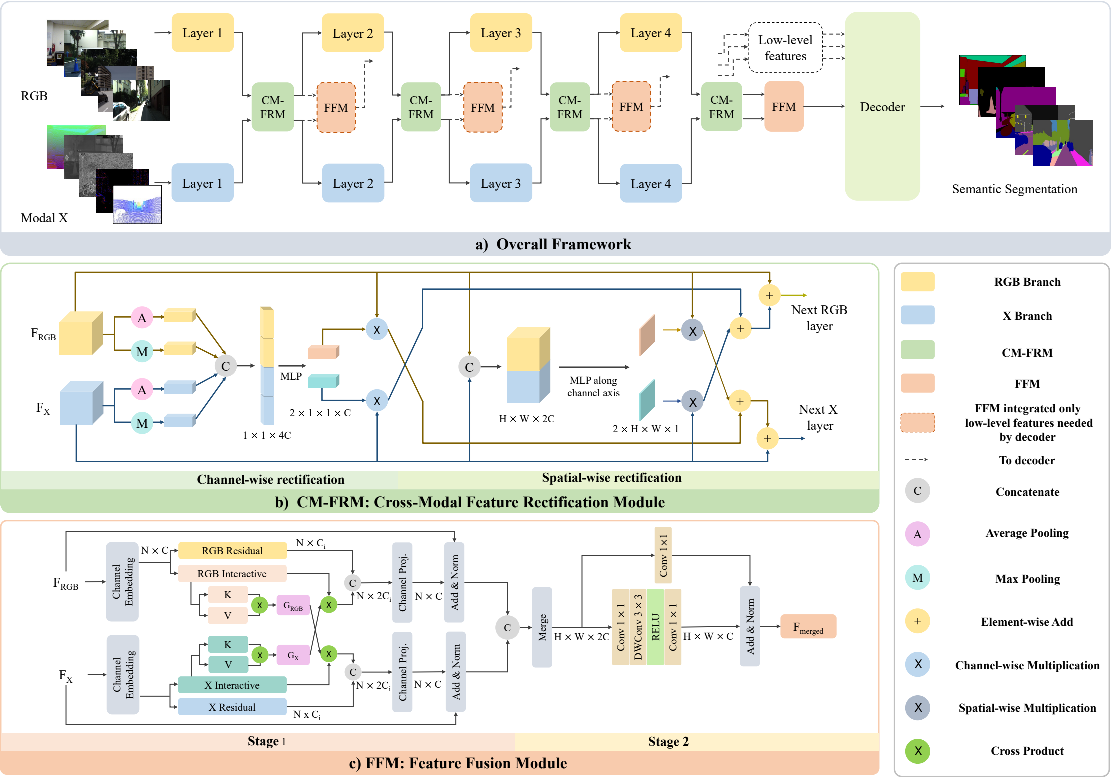


Figure 24: a) Overview of *CMX* for *RGB-X semantic segmentation*. The inputs are RGB and another modality (*e.g.*, Depth, Thermal, Polarization, Event, or LiDAR). b) Cross-Modal Feature Rectification Module (*CM-FRM*) with colored arrows as information flows of the two modalities. c) Feature Fusion Module (*FFM*) with two stages of information exchange and fusion.

4.1.2 *CMX Framework*

The overview of CMX is shown in Figure 24a. We use two parallel branches to extract features from RGB- and X-modal inputs, which can be RGB-Depth, -Thermal, -Polarization, -Event, -LiDAR data, *etc*. Specifically, our proposed framework for RGB-X semantic segmentation adopts a two-branch design to effectively extract features from both RGB- and X-modal inputs. The two branches involve the simultaneous processing of RGB- and X-modal data in a parallel but interactive manner, each of which is designed to capture the unique characteristics of the respective input modality.

4.1.2.1 *Cross-Modal Feature Rectification*

To perform feature rectification between parallel streams at each stage in feature extraction, we propose a novel *Cross-Modal Feature Rectification Module (CM-FRM)*, as shown in Figure 24b. CM-FRM processes features in two dimensions, including *channel-wise* and *spatial-wise* feature rectifications, which together offer a holistic calibration, enabling better multimodal feature extraction and interaction.

Channel-wise feature rectification. We embed bi-modal features $\mathrm{RGB}_{in} \in \mathbb{R}^{H \times W \times C}$ and $\mathrm{X}_{in} \in \mathbb{R}^{H \times W \times C}$ along the spatial axis into two attention vectors $W^{C}_{\mathrm{RGB}} \in \mathbb{R}^{C}$ and $W^{C}_{\mathrm{X}} \in \mathbb{R}^{C}$. Different from previous channel-wise attention methods [33, 36, 48], we apply both global max pooling and global average pooling to RGB_{in} and X_{in} along the channel dimension to retain more information. We concatenate the four resulted vectors, having $Y \in \mathbb{R}^{4C}$. Then, an MLP is applied, followed by a `sigmoid` function to obtain $W^C \in \mathbb{R}^{2C}$ from Y, which will be split into W^C_{RGB} and W^C_X:

$$W^C_{RGB}, W^C_X = \mathcal{F}_{split}(\sigma(\mathcal{F}_{mlp}(Y))), \tag{15}$$

where $\sigma(\cdot)$ denotes the `sigmoid` function. The channel-wise rectification is formed as:

$$\begin{aligned} RGB^C_{rec} &= W^C_X \circledast X_{in}, \\ X^C_{rec} &= W^C_{RGB} \circledast RGB_{in}, \end{aligned} \tag{16}$$

where $\circledast$ denotes channel-wise multiplication.

Spatial-wise feature rectification. The bi-modal inputs RGB_{in} and X_{in} will be concatenated and embedded into two spatial weight maps: $W^S_{RGB} \in \mathbb{R}^{H \times W}$ and $W^S_X \in \mathbb{R}^{H \times W}$. The embedding operation has two 1×1 convolution layers assembled with a RELU function. Afterward, a `Sigmoid` function is applied to obtain the embedded feature map $F \in \mathbb{R}^{H \times W \times 2}$, which is further split into two weight maps. The process to obtain the spatial weight maps is formulated as:

$$F = Conv_{1\times1}(RELU(Conv_{1\times1}(RGB_{in} \parallel X_{in}))), \tag{17}$$

$$W^S_{RGB}, W^S_X = \mathcal{F}_{split}(\sigma(F)). \tag{18}$$

Similar to channel-wise rectification, spatial-wise rectification is formulated as:

$$\begin{aligned} RGB^S_{rec} &= W^S_X * X_{in}, \\ X^S_{rec} &= W^S_{RGB} * RGB_{in}, \end{aligned} \tag{19}$$

where $*$ denotes spatial-wise multiplication. The whole rectified feature for both modalities RGB_{out} and X_{out} is organized as:

$$\begin{aligned} RGB_{out} &= RGB_{in} + \lambda_C RGB^C_{rec} + \lambda_S RGB^S_{rec}, \\ X_{out} &= X_{in} + \lambda_C X^C_{rec} + \lambda_S X^S_{rec}. \end{aligned} \tag{20}$$

λ_C and λ_S are two hyperparameters. We set them both as 0.5 as default. RGB_{out} and X_{out} are the rectified features after the comprehensive calibration, which will be sent into the next stage for feature fusion.

4.1.2.2 *Feature Fusion*

After obtaining multiple feature maps, we build a two-stage *Feature Fusion Module (FFM)* to enhance the information interaction and combination. As shown in Figure 24(c), in the information exchange stage (Stage 1), the two branches are still maintained, and a cross-attention mechanism is designed to globally exchange information between the two branches. In the fusion stage (Stage 2), the concatenated feature is transformed into the original size via a mixed channel embedding.

Information exchange stage. At this stage, the bi-modal features will exchange their information via a symmetric dual-path structure. For brevity, we take the X-modal path for illustration. We first flatten the input feature with size $\mathbb{R}^{H \times W \times C}$ to $\mathbb{R}^{N \times C}$, where

Dataset	Image	Event	Train/Val	Label	Resolution	Class
DDD17 [3]	Gray-scale	50Hz	15950/3890	pseudo	346 × 260	6
DSEC-Semantic [199]	Gray-scale	20Hz	8082/2809	pseudo	640 × 440	11
EventScape [64]	RGB	500Hz	122329/22493	synthetic	512 × 256	12

Table 8: Comparison of event-based semantic segmentation datasets.

$N{=}H{\times}W$. Afterward, a linear embedding is used to generate the residual vector X^{res} and interactive vector X^{inter} with the size $\mathbb{R}^{N\times C_i}$. We further put forward an efficient cross-attention mechanism [179] applied to these two interactive vectors from different modal paths, which will carry out sufficient information exchange across modalities. Specifically, the interactive vectors will be embedded into K and V for each head, and both sizes of them are $\mathbb{R}^{N\times C_{head}}$. The output is obtained by multiplying the interactive vector and the context vector from the other modality path, namely a cross-attention process, and it is depicted in the following equations:

$$G_{RGB} = K_{RGB}^{T} V_{RGB}, \quad G_{X} = K_{X}^{T} V_{X}, \tag{21}$$

$$U_{RGB} = X_{RGB}^{inter}\ \mathrm{SoftMax}(G_{X}), \quad U_{X} = X_{X}^{inter}\ \mathrm{SoftMax}(G_{RGB}). \tag{22}$$

Note that G denotes the global context vector, while U indicates the attended result. To realize the attention from different representation subspaces, we remain the multi-head mechanism, where the number of heads matches the transformer backbone. Then, the attended result vector U and the residual vector X^{res} are concatenated. Finally, we apply a second linear embedding and resize the feature to $\mathbb{R}^{H\times W\times C}$.

Fusion stage. In the second stage of FFM, *i.e.*, the fusion stage, we use a channel embedding to merge features from two paths, which is realized via $1{\times}1$ convolution layers. Further, inspired by Mix-FFN [238] and ConvMLP [104], we add one more depth-wise convolution layer $DWConv_{3\times3}$ to realize a skip-connected structure. In this way, the merged features with the size $\mathbb{R}^{H\times W\times 2C}$ are fused into the final output with the size of $\mathbb{R}^{H\times W\times C}$ for feature decoding.

4.1.3 *RGB-Event Semantic Segmentation Benchmark*

A large-scale multimodal RGB-Event semantic segmentation benchmark is not available. To fill this gap, we create an RGB-Event multimodal semantic segmentation benchmark[1] based on the EventScape dataset [64], which is originally designed for depth estimation. The comparison between three event-based semantic segmentation datasets is presented in Table 8. Unlike previous datasets using gray-scale images and pseudo labels, the RGB and the synthetic labels are available in our benchmark, which can provide more sufficient information and more precise annotations. To maintain data diversity from the original sequences generated by CARLA simulator [53], we select one frame from every 30 frames, obtaining 4077/749 images from 122329/22493 for training/evaluation. The images have a 512×256 resolution and are annotated with 12 semantic classes, including *Vehicle*, *Building*, *Wall*, *Vegetation*, *Road*, *Pole*, *RoadLines*, *Fences*, *Pedestrian*, *TrafficSign*, *Sidewalk*, and *TrafficLight*.

[1] RGB-Event: https://paperswithcode.com/sota/semantic-segmentation-on-eventscape.

(a) Results on NYU Depth V2 [185].

Method	mIoU (%)	Acc (%)
3DGNN [157]	43.1	-
Kong et al. [98]	44.5	72.1
LS-DeconvNet [41]	45.9	71.9
CFN [115]	47.7	-
ACNet [82]	48.3	-
RDF-101 [150]	49.1	75.6
SGNet [31]	51.1	76.8
ShapeConv [20]	51.3	76.4
NANet [272]	52.3	77.9
SA-Gate [36]	52.4	77.9
CMX (MiT-B2)	54.1	78.7
CMX (MiT-B2)*	**54.4**	**79.9**
CMX (MiT-B4)	56.0	79.6
CMX (MiT-B4)*	**56.3**	**79.9**
CMX (MiT-B5)	56.8	79.9
CMX (MiT-B5)*	**56.9**	**80.1**

(b) Results on Stanford2D3D [4].

Method	mIoU (%)	Acc (%)
Depth-aware CNN [220]	39.5	65.4
MMAF-Net-152 [59]	52.9	76.5
ShapeConv-101 [20]	60.6	**82.7**
CMX (MiT-B2)	61.2	82.3
CMX (MiT-B4)	**62.1**	82.6

(c) Results on SUN-RGBD [188].

Method	mIoU (%)	Acc (%)
3DGNN [157]	45.9	-
RDF-152 [150]	47.7	81.5
CFN [115]	48.1	-
D-CNN [220]	42.0	-
ACNet [82]	48.1	-
TCD [265]	49.5	83.1
SGNet [31]	48.6	82.0
SA-Gate [36]	49.4	82.5
NANet [272]	48.8	82.3
ShapeConv [20]	48.6	82.2
CMX (MiT-B2)*	49.7	82.8
CMX (MiT-B4)*	52.1	83.5
CMX (MiT-B5)*	**52.4**	**83.8**

(d) Results on ScanNetV2 *test* set [45].

Method	Modal	mIoU (%)
PSPNet [294]	RGB	47.5
AdapNet++ [209]	RGB	50.3
3DMV (2d-proj) [46]	RGB-D	49.8
FuseNet [73]	RGB-D	53.5
SSMA [209]	RGB-D	57.7
GRBNet [158]	RGB-D	59.2
MCA-Net [181]	RGB-D	59.5
DMMF [182]	RGB-D	59.7
CMX (MiT-B2)	RGB-D	**61.3**

(e) Results on Cityscapes *val* set [44].

Method	Modal	Backbone	mIoU (%)
SwiftNet [147]	RGB	ResNet-18	70.4
ESANet [176]	RGB	ResNet-50	79.2
GSCNN [200]	RGB	WideResNet-38	80.8
CCNet [85]	RGB	ResNet-101	81.3
DANet [61]	RGB	ResNet-101	81.5
ACFNet [271]	RGB	ResNet-101	81.5
SegFormer [238]	RGB	MiT-B2	81.0
SegFormer [238]	RGB	MiT-B4	82.3
RFNet [195]	RGB-D	ResNet-18	72.5
PADNet [242]	RGB-D	ResNet-50	76.1
Kong et al. [98]	RGB-D	ResNet-101	79.1
ESANet [176]	RGB-D	ResNet-50	80.0
SA-Gate [36]	RGB-D	ResNet-50	80.7
SA-Gate [36]	RGB-D	ResNet-101	81.7
AsymFusion [225]	RGB-D	Xception65	82.1
SSMA [209]	RGB-D	ResNet-50	82.2
CMX	RGB-D	MiT-B2	**81.6**
CMX	RGB-D	MiT-B4	**82.6**

Table 9: Results on five RGB-Depth datasets. * denotes multi-scale test.

4.1.4 *Experiments and Analysis*

4.1.4.1 *Results on RGB-Depth Datasets*

We first conduct experiments on RGB-D semantic segmentation datasets. The results are grouped in Table 9.

NYU Depth V2. The results on the NYU Depth V2 dataset are in Table 9a. It can be easily seen that our approach achieves leading scores. The proposed method with MiT-B2 already exceeds previous methods, attaining 54.4% in mIoU. Our CMX models based on MiT-B4 and -B5 further dramatically improve the mIoU to 56.3% and 56.9%, clearly standing out in front of all state-of-the-art approaches. The best CMX model even reaches superior results than recent strong pretraining-based methods [7, 66] like Omnivore [66] that uses images, videos, and single-view 3D data for supervision.

Stanford2D3D. In Table 9b, our CMX achieves state-of-the-art mIoU. B2-based CMX surpasses the previous best ShapeConv [20] based on ResNet-101 [74] and our model based on MiT-B4 further reaches mIoU to 62.1%. The results demonstrate the effectiveness and learning capacity of our approach on such a large RGB-D dataset.

SUN-RGBD. As presented in Table 9c, our method achieves leading performances on the SUN-RGBD dataset. Our interactive cross-modal fusion approach (Figure 22c) exceeds previous input fusion methods (Figure 22a), *e.g.*, SGNet [31] and ShapeConv [20], as well as feature fusion methods (Figure 22b), *e.g.*, ACNet [82] and SA-Gate [36]. In particular, with MiT-B4 and -B5, CMX elevates the mIoU to >52.0%. CMX is also better than multi-task methods like PAP [291] and TET [288].

ScanNetV2. We test our CMX model with MiT-B2 on the ScanNetV2 benchmark. As shown in Table 9d, it can be clearly seen that CMX outperforms RGB-only methods and achieves the top mIoU of 61.3% among the RGB-D methods. On the ScanNetV2 leaderboard, methods like BPNet [80] reach higher scores by using 3D supervision

Method	Modal	Unlabeled	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU
ERFNet [164]	RGB	96.7	67.1	56.2	34.3	30.6	9.4	0.0	0.1	30.5	36.1
DANet [61]	RGB	96.3	71.3	48.1	51.8	30.2	18.2	0.7	30.3	18.8	41.3
PSPNet [294]	RGB	96.8	74.8	61.3	50.2	38.4	15.8	0.0	33.2	44.4	46.1
HRNet [218]	RGB	98.0	86.9	67.3	59.2	35.3	23.1	1.7	46.6	47.3	51.7
SegFormer-B2 [238]	RGB	97.9	87.4	62.8	63.2	31.7	25.6	9.8	50.9	49.6	53.2
SegFormer-B4 [238]	RGB	98.0	88.9	64.0	62.8	38.1	25.9	6.9	50.8	57.7	54.8
MFNet [71]	RGB-T	96.9	65.9	58.9	42.9	29.9	9.9	0.0	25.2	27.7	39.7
SA-Gate [36]	RGB-T	96.8	73.8	59.2	51.3	38.4	19.3	0.0	24.5	48.8	45.8
Depth-aware CNN [220]	RGB-T	96.9	77.0	53.4	56.5	30.9	29.3	8.5	30.1	32.3	46.1
ACNet [82]	RGB-T	96.7	79.4	64.7	52.7	32.9	28.4	0.8	16.9	44.4	46.3
PSTNet [184]	RGB-T	97.0	76.8	52.6	55.3	29.6	25.1	**15.1**	39.4	45.0	48.4
RTFNet [196]	RGB-T	98.5	87.4	70.3	62.7	45.3	29.8	0.0	29.1	55.7	53.2
FuseSeg [198]	RGB-T	97.6	87.9	71.7	64.6	44.8	22.7	6.4	46.9	47.9	54.5
AFNet [243]	RGB-T	98.0	86.0	67.4	62.0	43.0	28.9	4.6	44.9	56.6	54.6
ABMDRNet [287]	RGB-T	**98.6**	84.8	69.6	60.3	45.1	33.1	5.1	47.4	50.0	54.8
FEANet [48]	RGB-T	98.3	87.8	71.1	61.1	46.5	22.1	6.6	**55.3**	48.9	55.3
DHFNet [18]	RGB-T	97.7	87.6	71.7	61.1	39.5	**42.4**	9.5	49.3	56.0	57.2
GMNet [305]	RGB-T	97.5	86.5	73.1	61.7	44.0	42.3	14.5	48.7	47.4	57.3
CMX (MiT-B2)	RGB-T	98.3	89.4	74.8	**64.7**	47.3	30.1	8.1	52.4	59.4	58.2
CMX (MiT-B4)	RGB-T	98.3	**90.1**	**75.2**	64.5	**50.2**	35.3	8.5	54.2	**60.6**	**59.7**

Table 10: Per-class results on MFNet dataset [71] for RGB-Thermal segmentation.

from point clouds to perform joint 2D- and 3D reasoning. In contrast, our method attains a competitively accurate performance by using purely 2D data and effectively leveraging the complementary information inside RGB-D modalities.

Cityscapes. Besides indoor RGB-D datasets, to study the generalizability to outdoor scenes, we assess the effectiveness of CMX on Cityscapes. As shown in Table 9e, we note that the improvement on the Cityscapes dataset is not as obvious as other datasets, because the performance of RGB-only models on this dataset shows a saturation trend. Compared with MiT-B2 (RGB), our RGB-D approach elevates the mIoU by 0.6%. Our approach based on MiT-B4 achieves a state-of-the-art score of 82.6%, outstripping all existing RGB-D methods by more than 0.4% in absolute mIoU values, verifying that CMX generalizes well to street scene understanding.

4.1.4.2 *Results on RGB-Thermal Dataset*

Comparison with the state-of-the-art. In Table 10, we compare our method against RGB-only models and multimodal methods using RGB-T inputs of MFNet dataset [71]. As unfolded, ACNet [82] and SA-Gate [36], carefully designed for RGB-Depth segmentation, perform less satisfactorily on RGB-T data, as they focus on feature extraction without sufficient feature interaction before fusion and thereby fail to generalize to other modalities. Depth-aware CNN [220], an input fusion method with modality-specific operator design, also does not yield high performance. In contrast, the proposed CMX strategy, enabling comprehensive interactions from various perspectives, generalizes smoothly in RGB-T semantic segmentation. It can be seen that our method based on MiT-B2 achieves mIoU of 58.2%, clearly outperforming the previous best RGB-T methods ABMDRNet [287], FEANet [48], and GMNet [305]. Our CMX with MiT-B4 further elevates state-of-the-art mIoU to 59.7%, widening the accuracy gap in contrast to existing methods. Moreover, it is worth pointing out that the improvements brought by our RGB-X approach compared with the RGB-only baselines are compelling, *i.e.*, +5.0% and +4.9% in mIoU for MiT-B2 and -B4 backbones, respectively. Our approach

Method	Modal	Building	Glass	Car	Road	Vegetation	Sky	Pedestrian	Bicycle	mIoU
SwiftNet [147]	RGB	83.0	73.4	91.6	96.7	94.5	84.7	36.1	82.5	80.3
SegFormer-B2 [238]	RGB	90.6	79.0	92.8	96.6	96.2	89.6	82.9	89.3	89.6
NLFNet [245]	RGB-P	85.4	77.1	93.5	97.7	93.2	85.9	56.9	85.5	84.4
EAFNet [235]	RGB-P	87.0	79.3	93.6	97.4	95.3	87.1	60.4	85.6	85.7
CMX (SegFormer-B2)	RGB-AoLP (Monochromatic)	**91.9**	87.0	95.6	98.2	96.7	89.0	84.9	92.0	91.8
CMX (SegFormer-B2)	RGB-AoLP (Trichromatic)	91.5	87.3	95.8	98.2	96.6	89.3	85.6	91.9	92.0
CMX (SegFormer-B4)	RGB-AoLP (Monochromatic)	91.8	**88.8**	**96.3**	**98.3**	96.7	89.1	86.3	92.3	92.4
CMX (SegFormer-B4)	RGB-AoLP (Trichromatic)	91.6	**88.8**	**96.3**	**98.3**	**96.8**	89.7	86.2	**92.8**	**92.6**
CMX (SegFormer-B2)	RGB-DoLP (Monochromatic)	91.4	87.6	96.0	98.2	96.6	89.1	87.1	92.3	92.1
CMX (SegFormer-B2)	RGB-DoLP (Trichromatic)	91.8	87.8	96.1	98.2	96.7	**89.4**	86.1	91.8	92.2
CMX (SegFormer-B4)	RGB-DoLP (Monochromatic)	91.8	88.6	**96.3**	**98.3**	96.7	89.4	86.0	92.1	92.4
CMX (SegFormer-B4)	RGB-DoLP (Trichromatic)	91.6	88.6	**96.3**	**98.3**	96.7	89.5	**86.4**	92.2	92.5

Table 11: Results on ZJU-RGB-P dataset [235] for RGB-Polarization segmentation.

overall achieves top scores on *car*, *person*, *bike*, *curve*, *car stop*, and *bump*. For *person* with infrared properties, our approach enjoys more than +11.0% gain in IoU, confirming the effectiveness of CMX in harvesting complementary cross-modal information.

4.1.4.3 *Results on RGB-Polarization Dataset*

Comparison with the state-of-the-art. Table 11 shows per-class accuracy of our approach compared to RGB-only [147, 238] and RGB-Polarization fusion methods [235, 245] on ZJU-RGB-P dataset [235]. Our unified CMX outperforms the previous best RGB-P method [235] by >6.0% in mIoU. We observe that the improvement on *pedestrian* is significant thanks to the capacity of the transformer backbone and our cross-modal fusion mechanisms. Compared to the RGB-only baseline with MiT-B2 [238]), the IoU improvements on classes with polarimetric characteristics are clear, such as *glass* (>8.0%) and *car* (>2.5%), further evidencing the generalizability of our cross-modal fusion solution in bridging RGB-P streams.

Analysis of polarization data representations. We study polarimetric data representations and the results displayed in Table 11 indicate that the Angle of Linear Polarization (AoLP) and the Degree of Linear Polarization (DoLP) representations both carry effective polarization information beneficial for semantic scene understanding, which is consistent with the finding in [235]. Besides, trichromatic representations are consistently better than monochromatic representations used in previous RGB-P segmentation works [235, 245]. This is expected as the trichromatic representation provides more detailed information, which should be leveraged to fully unlock the potential of trichromatic polarization cameras.

4.1.4.4 *Results on RGB-Event Dataset*

Comparison with the state-of-the-art. In Table 12, we benchmark more than 10 semantic segmentation methods, including RGB-only methods, CNN-based [30, 147, 155, 231] and transformer-based [125, 238, 281] methods, as well as multimodal methods [36, 195, 285]. In contrast, our models improve performance by mixing RGB-Event features, as seen in Table 12 and Figure 25. Our model using MiT-B4 reaches 64.28% in mIoU, towering over all other methods and setting the state-of-the-art on the RGB-

Method	Modal	Backbone	mIoU (%)	Pixel Acc. (%)
SwiftNet [147]	RGB	ResNet-18	36.67	83.46
Fast-SCNN [155]	RGB	Fast-SCNN	44.27	87.10
CGNet [231]	RGB	M3N21	44.75	87.13
Trans4Trans [281]	RGB	PVT-B2	51.86	89.03
Swin-s [125]	RGB	Swin-s	52.49	88.78
Swin-b [125]	RGB	Swin-b	53.31	89.21
DeepLabV3+ [30]	RGB	ResNet-101	53.65	89.92
SegFormer-B2 [238]	RGB	MiT-B2	58.69	91.21
SegFormer-B4 [238]	RGB	MiT-B4	59.86	91.61
RFNet [195]	RGB-E	ResNet-18	41.34	86.25
ISSAFE [285]	RGB-E	ResNet-18	43.61	86.83
SA-Gate [36]	RGB-E	ResNet-101	53.94	90.03
CMX (DeepLabV3+)	RGB-E	ResNet-101	54.91	89.67
CMX (Swin-s)	RGB-E	Swin-s	60.86	91.25
CMX (Swin-b)	RGB-E	Swin-b	61.21	91.61
CMX (SegFormer-B2)	RGB-E	MiT-B2	61.90	91.88
CMX (SegFormer-B4)	RGB-E	MiT-B4	**64.28**	**92.60**

Table 12: Results for RGB-Event segmentation.

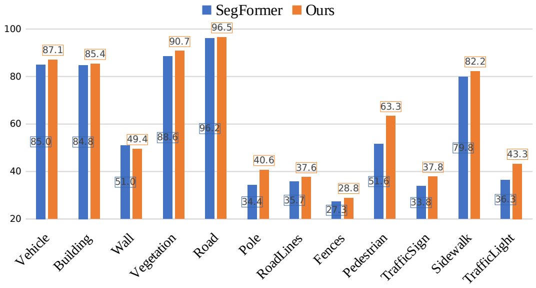


Figure 25: Per-class IoU results on RGB-Event benchmark.

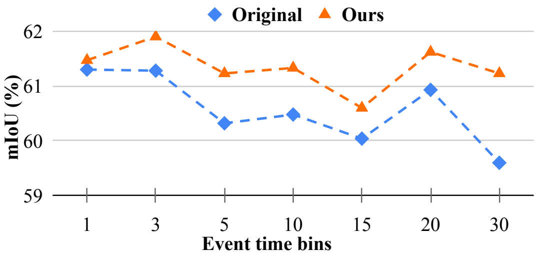


Figure 26: Analysis of event representations and time bins.

E benchmark. This further verifies the versatility of our solution for different multi-modal combinations. Figure 25 depicts a per-class accuracy comparison between the RGB baseline and our RGB-Event model with MiT-B2. With event data, the foreground objects are more accurately parsed by our RGB-E model, *e.g.*, *vehicle* (+2.1%), *pedestrian* (+11.7%), and traffic light (+7.0%).

Analysis of using different backbones. To verify that our unified method is effective with using different backbones, we compare CNN- and transformer-based backbones in the CMX framework. Specifically, in addition to MiT backbones, we experiment with DeepLabV3+ [30] and Swin transformer [125] backbones with UperNet [236] to construct CMX. Compared to the RGB-only DeepLabV3+, Swin-s, and Swin-b methods, CMX models achieve respective +1.26%, +8.37%, +7.90% gains in mIoU. The results show that our RGB-X solution consistently improves the segmentation performance, confirming that our unified framework is not strictly tied to a concrete backbone type, but can be flexibly deployed with CNN or transformer models, which helps to yield effective unified architecture for RGB-X semantic segmentation.

Analysis of event data representations. We study with different settings of event time bin B={1,3,5,10,15,20,30} based on our CMX fusion model with MiT-B2. Compared with the original event representation [64], our representation achieves consistent im-

Method	Backbone	mIoU (%)
HRFuser [14]	HRFormer-T	48.74
PMF [309]	SalsaNext	54.48
TokenFusion [224]	MiT-B2	54.55
TransFuser [156]	RegNetY-3.2GF	56.57
CMX	MiT-B2	**64.31**

Table 13: Results for RGB-LiDAR segmentation.

provements (in Figure 26) on different settings of event time bins, such as +1.63% of mIoU when B=30. In particular, it helps our CMX to obtain the highest mIoU of 61.90% in the setting of B=3. In B=1, embedding all events in a single time bin leads to dragging behind images of moving objects and being sub-optimal for feature fusion. In higher time bins, events produced in a short interval are dispersed to more bins, resulting in insufficient events in a single bin. These corroborate observations in [285, 286] and that the event representation B=3 is an effective time bin setting for RGB-E semantic segmentation with CMX.

4.1.4.5 *Results on RGB-LiDAR Dataset*

In Table 13, we compare CMX with other models dedicated to RGB-LiDAR data fusion, including PMF [309] and TransFuser [156]. These two methods achieve respective 54.48% and 56.57% in mIoU. Besides, other general multimodal fusion methods, *e.g.*, HRFuser [14] and TokenFusion [224], are included for comparison. In contrast, our CMX obtains the best performance with 64.31% in mIoU, having a +9.76% gain compared with TokenFusion which is also based on MiT-B2. The sufficient improvement proves the advantage of using a symmetric dual-stream architecture in modal fusion and the effectiveness of our proposed cross-modal rectification and fusion methods.

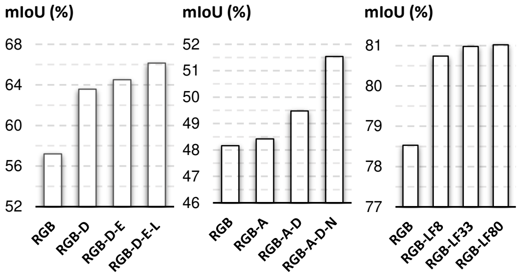
4.2 ARBITRARY-MODAL SEMANTIC SEGMENTATION

This section is based on our work published in *CVPR 2023* [279].

4.2.1 *Arbitrary-Modal Data Fusion*

With the explosion of modular sensors, multimodal fusion for semantic segmentation has progressed rapidly recently [20, 36, 278] and in turn has stirred growing interest to assemble more and more sensors to reach higher and higher segmentation accuracy aside from more robust scene understanding. However, most works [82, 233, 309] and multimodal benchmarks [71, 185, 285] focus on specific sensor pairs, which lack behind the current trend of fusing more and more modalities [14, 224], *i.e.*, progressing towards Arbitrary-Modal Semantic Segmentation (AMSS).

When examining AMSS, two observations become evident: *(1) An increasing amount of modalities should provide more diverse complementary information, monotonically increasing segmentation accuracy.* This is directly supported by our results when incrementally adding and fusing modalities as illustrated in Figure 27a (RGB-Depth-Event-LiDAR), Figure 27b (RGB-AoLP-DoLP-NIR), and Figure 27c when adding up to 80 sub-aperture light-field modalities (RGB-LF8/-LF33/-LF80). Unfortunately, this great potential cannot be uncovered by previous cross-modal fusion methods [32, 235, 305], which follow designs for pre-defined modality combinations. *(2) The cooperation of multiple sensors is expected to effectively combat individual sensor failures.* Most of the existing works [210, 225, 232] are built on the assumption that each modality is always accurate. Under partial sensor faults, which are common in real-life robotic systems, *e.g*. LiDAR Jitter, fusing misaligned sensing data might even degrade the segmentation performance, as depicted with CMX [278] and HRFuser [14] in Figure 28. These two critical observations remain to a large extent neglected.



(a) RGB-D-E-L (b) RGB-A-D-N (c) RGB-Light Field

Figure 27: Arbitrary-modal segmentation results of CMNeXt on three datasets.

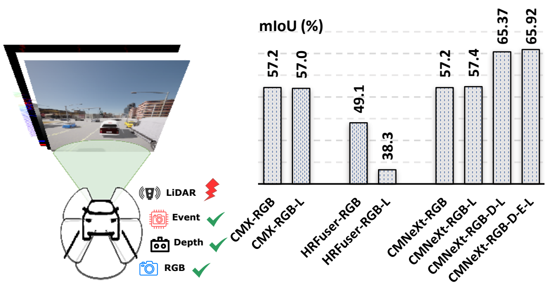
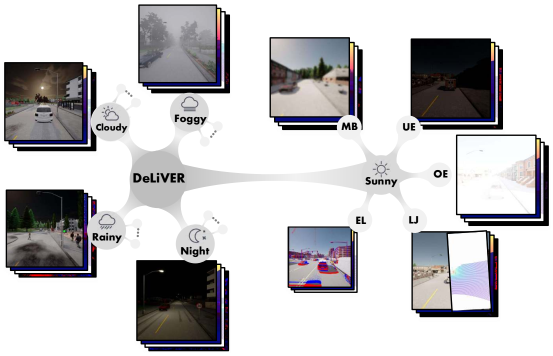


Figure 28: Comparison in sensor failure (*i.e.*, LiDAR Jitter) on the DELIVER dataset.

4.2.2 *The new DELIVER Dataset*

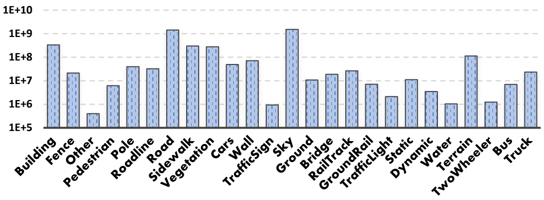
To address these challenges, we create a benchmark based on the CARLA simulator [53], with **De**pth, **Li**DAR, **V**iews, **E**vents, and **R**GB images: the **DELIVER** multimodal dataset. It features severe weather conditions and five sensor failure modes to exploit complementary modalities and resolve partial sensor outages.



(a) Structure and samples of four adverse conditions and five failure cases.

Split	Cloudy	Foggy	Night	Rainy	Sunny	Normal	Corner	Total
Train	794	795	797	799	798	2585	1398	3983
Val	398	400	410	398	399	1298	707	2005
Test	379	379	379	380	380	1198	699	1897
Front-view	1571	1574	1586	1577	1577	5081	2804	7885
All six views	9426	9444	9516	9462	9462	30486	16824	47310

(b) Statistic of different data splits and views.



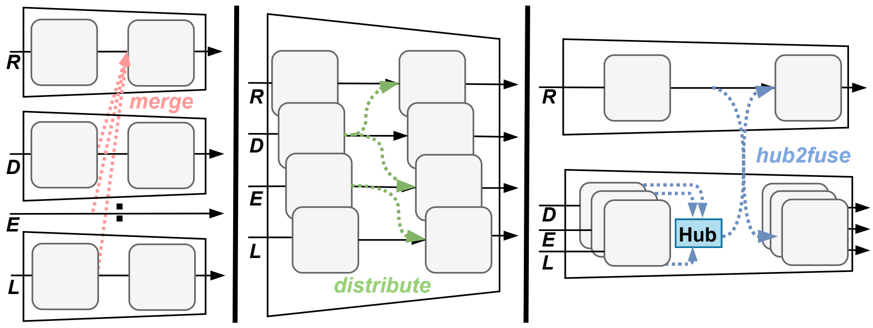
(c) Distribution of 25 semantic classes in logarithmic scaling.

Figure 29: DeLiVER multimodal dataset including (a) four adverse conditions out of five conditions (*i.e.*, *cloudy*, *foggy*, *night-time*, *rainy* and *sunny*). Apart from normal cases, each condition has five corner cases (*i.e.*, **MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution). Each sample has six views. Each view has four modalities and two labels (*i.e.*, semantic and instance). (b) is the data statistics. (c) is the data distribution of 25 semantic classes.

Sensor settings and modalities. As presented in Figure 29, we spent the effort to create a large-scale multimodal segmentation dataset DeLiVER, which provides six mutually orthogonal views (*i.e.*, *front, rear, left, right, up, down*) of the same spatial viewpoint, *i.e.*, a complete frame of data is encoded in the format of a panoramic cubemap. The Field-of-View (FoV) of each view is 91°×91° and the image resolution is 1042×1042. All Depth, Views, and Event sensors use the same camera settings when the sensor is working properly. According to the characteristics of recent LiDAR sensors [63], we further customize a 64 vertical channels virtual semantic LiDAR sensor to generate a point cloud of 1,728,000 points per second with a FoV of 360°× (-30°∼10°) and a range of 100 meters, so as to collect relatively dense LiDAR data.

Adverse conditions and corner cases. In addition to the multimodal setup, DeLiVER provides cases in two-fold, including four environmental conditions and five partial sensor failure cases (Figure 29a). For environmental conditions, we consider *cloudy*, *foggy*, *night*, and *rainy* weather conditions other than only *sunny* days. The environmental conditions will cause variations in the position and illumination of the sun, atmospheric diffuse reflections, precipitation, and shading of the scene, introducing challenges for robust perception. For sensor failure cases, we consider Motion Blur (MB), Over-Exposure (OE), and Under-Exposure (UE) common for RGB cameras. LiDAR failures usually manifest as along-axis LiDAR-Jitter (LJ) due to fixation issues or rotational axis eccentricity, thus we add random angular jitters in the range of [-1°, 1°] and position jitters of [-1cm, 1cm] to the three axial directions of the LiDAR sensor. Due to the circuit design, the resolution of the currently-used event sensors is limited [62]. Thus, we customize an Event Low-resolution (EL) scenario with 0.25× resolution for the event camera to simulate actual devices.

Statistics and annotations. Including six views, DeLiVER has totally 47,310 frames (Figure 29b) with the size of 1042×1042. The 7,885 front-view samples are divided into 3,983/2,005/1,897 for training/validation/testing, respectively, each of which contains



(a) Separate (b) Joint (c) Asymmetric

Figure 30: Comparison of fusion paradigms, such as (a) merging with separate branches [14], (b) distributing with a joint branch [224], and (c) our hub2fuse with asymmetric branches.

two types of annotations (*i.e.*, semantic and instance labels). Note that, we mainly discuss the front view and the semantic segmentation task in this work, while other views and instance segmentation will be future works. To improve the class diversity of annotations (25 classes as in Figure 29c), we modify and remap the semantic labels in the source code. Specifically, the *Vehicles* class is subdivided into four fine-grained categories: *Cars*, *TwoWheeler*, *Bus*, and *Truck* for both the semantic camera and the semantic LiDAR, making DeLiVER compatible with popular segmentation datasets. More details of the dataset are presented in Appendix A.1.

4.2.3 *CMNeXt Model*

We present the arbitrary-modal segmentation model *CMNeXt*. Without increasing the computation overhead substantially when adding more modalities, CMNeXt incorporates a novel *Hub2Fuse* paradigm (Figure 30c). Unlike relying on separate branches (Figure 30a) which tend to be computationally costly or using a single joint branch (Figure 30b) which often discards valuable information, CMNeXt is asymmetric with two branches, one for RGB and another for diverse supplementary modalities.

The key challenge lies in designing the two branches to pick up multimodal cues. Specifically, at the *hub* step of *Hub2Fuse*, to gather useful complementary information from auxiliary modalities, we design a Self-Query Hub (SQ-Hub), which dynamically selects informative features from all modality-sources before fusion with the RGB branch. Another great benefit of SQ-Hub is the ease of extending it to an arbitrary number of modalities, at negligible parameters increase (~0.01M per modality). At the *fusion* step, fusing sparse modalities such as LiDAR or Event data can be difficult to handle for joint branch architectures without explicit fusion such as TokenFusion [224]. To circumvent this issue and make best use of both dense and sparse modalities, we leverage cross-fusion modules [278] and couple them with our proposed Parallel Pooling Mixer (PPX) which efficiently and flexibly harvests the most discriminative cues from any auxiliary modality. These design choices come together in our CMNeXt architecture, which paves the way for AMSS (Figure 27). By carefully putting together alternative modalities, CMNeXt can overcome individual sensor failures and enhances segmentation robustness (Figure 28).

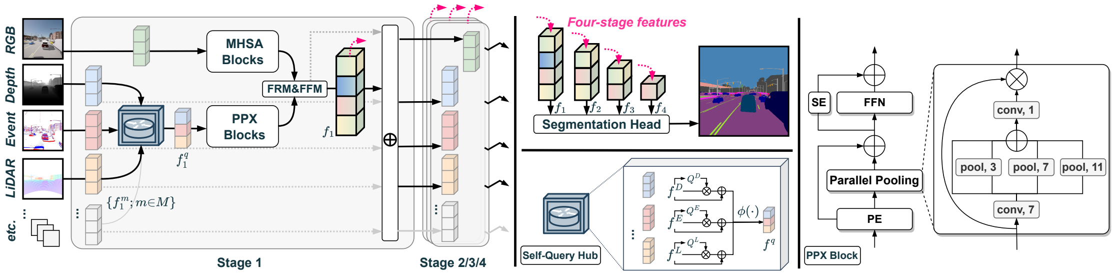


Figure 31: CMNeXt architecture in Hub2Fuse paradigm and asymmetric branches, having *e.g.* Multi-Head Self-Attention (MHSA) [238] blocks in the RGB branch and our Parallel Pooling Mixer (PPX) blocks in the accompanying branch. At the *hub* step, the Self-Query Hub selects informative features from the supplementary modalities. At the *fusion* step, the feature rectification module (FRM) and feature fusion module (FFM) [278] are used for feature fusion. Between stages, features of each modality are restored via adding the fused feature. The four-stage fused features are forwarded to the segmentation head for the final prediction.

In Figure 31, our CMNeXt has an encoder-decoder architecture. Built on the assumption that the RGB representation is essential for semantic segmentation, the two branches correspond to the primary branch for RGB and the secondary branch for other modalities, respectively. The four-stage structure follows most of previous CNN/Transformer models [61, 222, 238, 294] to extract pyramidal features. Note that, Figure 31 details only the first of the four stages for brevity. For the consistency of modal representations, we preprocess LiDAR and Event data as image-like representations following [285, 309]. The RGB image $\boldsymbol{I}_{\mathrm{RGB}}{\in}H{\times}W{\times}3$ is gradually processed by Multi-Head Self-Attention (MHSA) blocks [238], whereas the images of the other M modalities $\boldsymbol{I}_{M}{\in}H{\times}W{\times}3{\times}M$ by Parallel Pooling Mixer (PPX) blocks. After four stages, there are $M{+}1$ sets of four-stage feature maps $\boldsymbol{f}_l^m{\in}\{\boldsymbol{f}_1^m, \boldsymbol{f}_2^m, \boldsymbol{f}_3^m, \boldsymbol{f}_4^m\}$, $m{\in}[1, M{+}1]$. In the l^{th} stage, the block number of each branch is $b_l{\in}\{4, 8, 16, 32\}$, the stride is $s_l{\in}\{4, 8, 16, 32\}$, and the channel dimension is $C_l{\in}\{64, 128, 320, 512\}$. Inside each stage, $M{+}1$ features are processed in the *Hub2Fuse* paradigm: At the *hub* step, M feature maps will be merged into one feature $\boldsymbol{f}^q$ via the proposed Self-Query Hub. At the *fusion* step, the merged feature $\boldsymbol{f}^q$ will be further fused with RGB feature by the cross-modal Feature Rectification Module (FRM) [278] and Feature Fusion Module (FFM) [278], termed as $\boldsymbol{f}$. These two modules enable better multimodal feature fusion and interaction, and are crucial when fusing RGB with sparse features, which will be shown in our experiments. Between stages, $M{+}1$ feature maps will be restored via adding the fused feature $\boldsymbol{f}$, respectively. After the encoder, the four-stage features $\boldsymbol{f}_l{\in}\{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_4\}$ will be forwarded to the decoder for the segmentation prediction. We use the MLP decoder [238] as the segmentation head.

4.2.4 *Self-Query Hub*

To perform arbitrary-modal fusion, the Self-Query Hub (SQ-Hub) is a crucial design to select the informative features of supplementary modalities before fusing with the RGB feature. As shown in Figure 31, given a set of M supplementary features

$\{\boldsymbol{f}^{\mathrm{m}}|\mathrm{m}{\in}[1,M], \boldsymbol{f}^{\mathrm{m}}{\in}H{\times}W{\times}C\}$, a Self-Query module is applied to calculate the informative score mask $Q^{\mathrm{m}}{\in}H{\times}W$ of each feature $\boldsymbol{f}^{\mathrm{m}}$, as in Eq. (23) and (24).

$$\begin{aligned} \hat{\boldsymbol{f}}^{\mathrm{m}} &= \text{DW-Conv}_{3\times3}(C,C)(\boldsymbol{f}^{\mathrm{m}}), && (23)\\ Q^{\mathrm{m}} &= \mathtt{Sigmoid}(\text{Conv}(C,1)(\hat{\boldsymbol{f}}^{\mathrm{m}})), && (24)\end{aligned}$$

where the $\text{DW-Conv}_{3\times3}(C_{in}, C_{out})(\cdot)$ means a Depth-Wise convolution layer with a kernel size of $3{\times}3$. After obtaining M score masks through M respective self-query modules, a cross-modal comparison is conducted between M features $\{\boldsymbol{f}^{\mathrm{m}}|\mathrm{m}{\in}[1,M]\}$. That is, each patch p^{q} of the merged feature map $\boldsymbol{f}^{q}$ will be filled by the patch p^{m} of $\{\boldsymbol{f}^{\mathrm{m}}|\mathrm{m}{\in}[1,M]\}$ with the highest score, *i.e.*, the most effective patch among M modalities. It can be formalized as:

$$\begin{aligned} \boldsymbol{f}^{q} &= \{p^{q}|p^{q}{\in}H{\times}W\}\\ &= \phi(\{\boldsymbol{f}^{\mathrm{m}}{+}Q^{\mathrm{m}}{\cdot}\hat{\boldsymbol{f}}^{\mathrm{m}}|\mathrm{m}{\in}[1,M]\}) && (25)\\ &= \phi(\{p^{\mathrm{m}}|p^{\mathrm{m}}{\in}H{\times}W, \mathrm{m}{\in}[1,M]\}),\end{aligned}$$

where $\phi(\cdot)$ is an operation to select the maximum p^{m} from $\{\boldsymbol{f}^{\mathrm{m}}{+}Q^{\mathrm{m}}{\cdot}\hat{\boldsymbol{f}}^{\mathrm{m}}|\mathrm{m}{\in}[1,M]\}$. Then, the merged feature $\boldsymbol{f}^{q}$ is forwarded to the Parallel Pooling Mixer (PPX).

4.2.5 *Parallel Pooling Mixer*

Another crucial design in CMNeXt is the Parallel Pooling Mixer (Figure 31), which is proposed to efficiently and flexibly harvest discriminative cues from arbitrary-modal complements in the aforementioned SQ-Hub. Given the merged feature map $\boldsymbol{f}^{q}{\in}H{\times}W{\times}C$ from SQ-Hub, a $7{\times}7$ DW-Conv layer is applied to aggregate local information. The three parallel pooling layers are for capturing multi-scale modal features, which will be summed with the residual one and mixed by a $1{\times}1$ convolution. Then, a `Sigmoid` function is used to calculate the attention for weighting. The first part of PPX can be written as:

$$\begin{aligned} \hat{\boldsymbol{f}}^{q} &= \text{DW-Conv}_{7\times7}(C,C)(\boldsymbol{f}^{q}), && (26)\\ \hat{\boldsymbol{f}}^{q} &:= \sum_{k\in\{3,7,11\}} \text{Pool}_{k\times k}(\hat{\boldsymbol{f}}^{q}) + \hat{\boldsymbol{f}}^{q}, && (27)\\ \boldsymbol{w} &= \mathtt{Sigmoid}(\text{Conv}_{1\times1}(C,C)(\hat{\boldsymbol{f}}^{q})), && (28)\\ \boldsymbol{f}^{w} &= \boldsymbol{w}{\cdot}\boldsymbol{f}^{q} + \boldsymbol{f}^{q}. && (29)\end{aligned}$$

Previous methods [36, 82] show that channel information is crucial. Inspired by this, we apply a Squeeze-and-Excitation (SE) module [78] in the mixing part of PPX. This structure is crucial since some channels of certain modalities do capture more significant information than others. It can further engage more spatially-holistic knowledge in the channels of the cross-modal complements in SQ-Hub. Thus, the weighted feature $\boldsymbol{f}^{w}$ is passed to a Feed-Forward Network (FFN) and a SE module [78] for enhancing the channel information. The second part of PPX can be written as:

$$\hat{\boldsymbol{f}^{w}} = \text{FFN}(C,C)(\boldsymbol{f}^{w}) + \text{SE}(\boldsymbol{f}^{w}). \qquad (30)$$

(a) Results on KITTI-360 and DeLiVER datasets.

Method	Modal	Backbone	KITTI-360	DeLiVER
HRFuser [14]	RGB	HRFormer-T	53.20	47.95
SegFormer [238]	RGB	MiT-B2	67.04	57.20
HRFuser [14]	RGB-Depth	HRFormer-T	49.32	51.88
TokenFusion [224]	RGB-Depth	MiT-B2	57.44	60.25
CMX [278]	RGB-Depth	MiT-B2	64.43	62.67
CMNeXt	RGB-Depth	MiT-B2	65.09	63.58
HRFuser [14]	RGB-Event	HRFormer-T	44.85	42.22
TokenFusion [224]	RGB-Event	MiT-B2	55.97	45.63
CMX [278]	RGB-Event	MiT-B2	64.03	56.52
CMNeXt	RGB-Event	MiT-B2	66.13	57.48
HRFuser [14]	RGB-LiDAR	HRFormer-T	48.74	43.13
TokenFusion [224]	RGB-LiDAR	MiT-B2	54.55	53.01
CMX [278]	RGB-LiDAR	MiT-B2	64.31	56.37
CMNeXt	RGB-LiDAR	MiT-B2	65.26	58.04
HRFuser [14]	RGB-D-Event	HRFormer-T	50.21	51.83
CMNeXt	RGB-D-Event	MiT-B2	67.73	64.44
HRFuser [14]	RGB-D-LiDAR	HRFormer-T	52.61	52.72
CMNeXt	RGB-D-LiDAR	MiT-B2	66.55	65.50
HRFuser [14]	RGB-D-E-Li	HRFormer-T	52.76	52.97
CMNeXt	RGB-D-E-Li	MiT-B2	**67.84**	**66.30**

(b) Results on MFNet.

Method	Modal	mIoU
SwinT [125]	RGB	49.0
SegFormer [238]	RGB	52.0
ACNet [82]	RGB-T	46.3
FuseSeg [197]	RGB-T	54.5
ABMDRNet [287]	RGB-T	54.8
LASNet [101]	RGB-T	54.9
FEANet [48]	RGB-T	55.3
MFTNet [302]	RGB-T	57.3
GMNet [305]	RGB-T	57.3
DooDLeNet [60]	RGB-T	57.3
CMX (MiT-B2) [278]	RGB-T	58.2
CMX (MiT-B4) [278]	RGB-T	59.7
CMNeXt (MiT-B4)	RGB-T	**59.9**

(c) Results on NYU Depth V2.

Method	mIoU
ACNet [82]	48.3
SGNet [32]	51.1
ShapeConv [20]	51.3
NANet [273]	52.3
SA-Gate [36]	52.4
PGDENet [306]	53.7
TokenFusion [224]	54.2
TransD-Fusion [232]	55.5
MultiMAE [7]	56.0
Omnivore [66]	56.8
CMX (MiT-B4) [278]	56.3
CMX (MiT-B5) [278]	**56.9**
CMNeXt (MiT-B4)	**56.9**

(d) Results on UrbanLF-Real and -Syn.

Method	Modal	Real	Syn
PSPNet [294]	RGB	76.34	75.78
OCR [262]	RGB	78.60	79.36
SegFormer [238] (B4)	RGB	82.20	78.53
DAVSS [308]	Video	75.91	74.27
TMANet [215]	Video	77.14	76.41
ESANet [176]	RGB-D	*n.a.*	79.43
SA-Gate [36]	RGB-D	*n.a.*	79.53
PSPNet-LF [180]	RGB-LF33	78.10	77.88
OCR-LF [180]	RGB-LF33	79.32	80.43
CMNeXt (MiT-B4)	RGB-LF8	**83.22**	80.74
CMNeXt (MiT-B4)	RGB-LF33	82.62	80.98
CMNeXt (MiT-B4)	RGB-LF80	83.11	**81.02**

(e) Results on MCubeS.

Method	Modal	mIoU
DRConv [27]	RGB-A-D-N	34.63
DDF [303]	RGB-A-D-N	36.16
TransFuser [156]	RGB-A-D-N	37.66
MMTM [92]	RGB-A-D-N	39.71
FuseNet [73]	RGB-A-D-N	40.58
MCubeSNet [113]	RGB	33.70
CMNeXt (MiT-B2)	RGB	48.16
MCubeSNet [113]	RGB-A	39.10
CMNeXt (MiT-B2)	RGB-A	48.42
MCubeSNet [113]	RGB-A-D	42.00
CMNeXt (MiT-B2)	RGB-A-D	49.48
MCubeSNet [113]	RGB-A-D-N	42.86
CMNeXt (MiT-B2)	RGB-A-D-N	**51.54**

Table 14: Results on six multimodal semantic segmentation datasets. The KITTI-360 [114] and DeLiVER have four modalities. The MFNet [71] and NYU Depth V2 [185] are dual-modal with respective RGB-Thermal and RGB-Depth modalities. The UrbanLF [180] has 81 sub-aperture light-filed images. The MCubeS dataset [113] is quad-modal.

After the PPX block, $f^{\hat{w}}$ is fused with RGB feature to form the final fused feature $f_l \in \{f_1, f_2, f_3, f_4\}$ by using FRM&FFM modules [278], as shown in Figure 31.

Compared with convolution-based MSCA [70], pooling-based MetaFormer [261], fully-attentional FAN [300], our PPX includes two advances: (1) parallel pooling layers for efficient weighting in the attention part; (2) channel-wise enhancement in the feature mixing part. Both characteristics of the PPX block help in highlighting the cross-modal fused feature spatial- and channel-wise, respectively.

4.2.6 *Experiments and Analysis*

To verify the efficacy of our proposed CMNeXt framework, we conduct extensive experiments on six multimodal segmentation datasets. The results and comparisons against the state-of-the-art are shown in Table 14.

Results on DeLiVER. Table 14a summarizes the extensive comparisons between our CMNeXt and other recent methods on DeLiVER dataset. Overall, CMNeXt sets the state of the art on the fusion of two to four modalities. While fusing RGB with Depth, Event, and LiDAR, the bi-modal CMNeXt yields sufficient improvements, compared to HRFuser [14] and TokenFusion [224]. This demonstrates the superiority of our *Hub2Fuse* paradigm over the *seperate* and *joint* branch paradigm (Figure 30a and 30b), especially when fusing sparse modalities, *i.e.*, Event and LiDAR. From RGB-only to gradually fusing Depth, Events, and LiDAR, the mIoU scores of CMNeXt are gradually increased (57.20%→63.58%→64.44%→66.30%), showing the advance of arbitrary-modal fusion for segmentation. Thanks to the complementary features from other modalities, our quad-modal CMNeXt outperforms the RGB-only baseline SegFormer [238] by a significant margin of +9.10%.

Results on KITTI-360. In Table 14a, we found that most of the multimodal fusion methods on KITTI-360 [114] did not bring the expected high improvement. There are two conjectures: The samples are collected in suburbs and are composed of video sequences, resulting in insufficient scene diversity; The depth- and event data are generated from RGB sequences, resulting in limited modal differences. Thus, the segmentation output relies on the RGB segmentation, and adding modalities might be redundant. Nonetheless, our quad-modal CMNeXt achieves a +0.80% gain compared to the RGB-only baseline [238]. Besides, our bi-modal CMNeXt performs superior to CMX [278] by +1.56% to +2.85%. When fusing three to four modalities, CMNeXt has respective +17.52%, +13.94%, and +15.08% gains compared to HRFuser [14].

RGB-T and RGB-D segmentation. As shown in Table 14b and 14c, we further conduct experiments on bi-modal datasets, MFNet [71] and NYU Depth V2 [185], which comprise dense thermal and depth data as supplementary information. Our CMNeXt achieves the state of the art on both datasets. Using MiT-B4 [238], CMNeXt outperforms CMX with +0.2% on MFNet. Besides, on the NYU Depth V2 dataset, it is comparable to CMX with MiT-B5. It proves the benefits of our PPX block in CMNeXt over the Multi-Head Self-Attention (MHSA) block used by CMX.

Light field semantic segmentation. Towards arbitrary-modal fusion for semantic segmentation, we apply CMNeXt on the UrbanLF dataset [180], in which each sample is composed of 81 sub-aperture light field modalities. As shown in Table 14d, CMNeXt surpasses the previous state of the art, OCR-LF [180], in both real-world and synthetic scenes, even with fewer modalities (33→8). Due to the similarity between modalities in this dataset, it is challenging to extract diverse complementary features. Nonetheless, by fusing up to 80 light field images, CMNeXt reaches respective 83.11% and 81.02% in mIoU on real and synthetic sets.

Multimodal material segmentation. To verify multimodal fusion in material recognition, we conduct experiments on the quad-modal MCubeS dataset [113]. As shown in Table 14e, our quad-modal CMNeXt exceeds other quad-modal models and attains the top performance of 51.54%, with a significant increase 8.68% over MCubeSNet [113]. In addition, CMNeXt has incremental improvements when gradually adding AoLP, DoLP,

Model-modality	**#Params(M)**	**GFLOPs**	**Cloudy**	**Foggy**	**Night**	**Rainy**	**Sunny**	**MB**	**OE**	**UE**	**LJ**	**EL**	**Mean**
HRFuser-RGB	29.89	217.5	49.26	48.64	42.57	50.61	50.47	48.33	35.13	26.86	49.06	49.88	47.95
SegFormer-RGB	25.79	38.93	59.99	57.30	50.45	58.69	60.21	57.28	56.64	37.44	57.17	59.12	57.20
TokenFusion-RGB-D	26.01	54.96	50.92	52.02	43.37	50.70	52.21	49.22	46.22	36.39	49.58	49.17	49.86
CMX-RGB-D	66.57	65.68	63.70	62.77	60.74	62.37	63.14	59.50	60.14	55.84	62.65	63.26	62.66
HRFuser-RGB-D	30.46	223.0	54.80	51.48	49.51	51.55	52.12	50.92	41.51	44.00	54.10	52.52	51.88
HRFuser-RGB-D-E	31.04 (+0.57)	229.0 (+6.00)	54.04	50.83	50.88	51.13	52.61	49.32	41.75	47.89	54.65	52.33	51.83
HRFuser-RGB-D-E-L	31.61 (+0.57)	235.0 (+6.00)	56.20	52.39	49.85	52.53	54.02	49.44	46.31	46.92	53.94	52.72	52.97
CMNeXt-RGB-D	58.69	62.94	67.21	62.79	61.64	62.95	65.26	61.00	64.64	58.71	64.32	63.35	63.58
CMNeXt-RGB-D-E	58.72 (+0.03)	64.19 (+1.25)	68.28	63.28	62.64	63.01	66.06	62.58	64.44	58.73	65.37	65.80	64.44
CMNeXt-RGB-D-E-L	58.73 (+0.01)	65.42 (+1.23)	68.70	65.67	62.46	67.50	66.57	62.91	64.59	60.00	65.92	65.48	66.30
w.r.t. SegFormer-RGB			(+8.71)	(+8.37)	(+12.01)	(+8.81)	(+6.36)	(+5.63)	(+7.95)	(+22.56)	(+8.75)	(+6.36)	(+9.10)

Table 15: Results on adverse conditions of DeLiVER. Sensor failure cases are **MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution. The number of parameters (#Params) and GFLOPs are counted in 512×512.

and NIR modalities. The results on multimodal material segmentation are consistent with the ones of arbitrary-modal segmentation on our DeLiVER dataset.

Analysis in adverse weather conditions. In Table 15, we compare CMNeXt against mainstream fusion paradigms in different conditions including adverse weather situations and partial sensor failure scenarios. It can be seen that despite being efficient, TokenFusion [224] suffers in these conditions as effective information is discarded in their token replacement. Due to the proposed SQ-Hub for selecting effective features, CMNeXt significantly improves the performance compared to the previous CMX [278] and HRFuser [14]. When fusing more modalities, HRFuser tends to induce much more overhead (+6.00 GFLOPs when adding a branch), whereas CMNeXt brings great mIoU gains at only slight computation increase (<1.30 GFLOPs). Compared to the RGB baseline, the RGB-D-E-L CMNeXt improves the accuracy by 9.10% on average, in particular for the nighttime (+12.01%) and the rainy (+8.81%) scenarios.

Analysis in sensor failure cases. In the Event Low-resolution (**EL**) case of Table 15, from the fusion of RGB-D to RGB-D-E, the accuracy of HRFuser [14] is degraded, however, the one of CMNeXt is improved (63.35%→66.11%). This is also observed in the case of LiDAR Jitter (**LJ**), where the performance of CMNeXt is increased (65.37%→65.92%) by fusing from D-E to D-E-L. These results demonstrate the ability of CMNeXt to combat sensor failures, thanks to SQ-Hub for selecting informative features. Compared to the RGB baseline, CMNeXt obtains a +22.56% gain in the Under-Exposure (**UE**) case.

Visualization of arbitrary-modal segmentation. In Figure 32, we show semantic segmentation results of our CMNeXt against the RGB-only SegFormer [238] and the RGB-X CMX [278]. It can be seen that in the dark night with under-exposure, the RGB-only SegFormer hardly segments the close vehicle, while the RGB-D CMNeXt clearly outperforms CMX. Our RGB-D-E-L CMNeXt further enhances the performance and yields more complete segmentation. In the partial sensor failure scenario with LiDAR jitter, CMX produces unsatisfactory rainy scene parsing results. Our RGB-LiDAR model is barely affected by the sensing data mis-alignment and the quad-modal CMNeXt further robustifies the full scene segmentation.

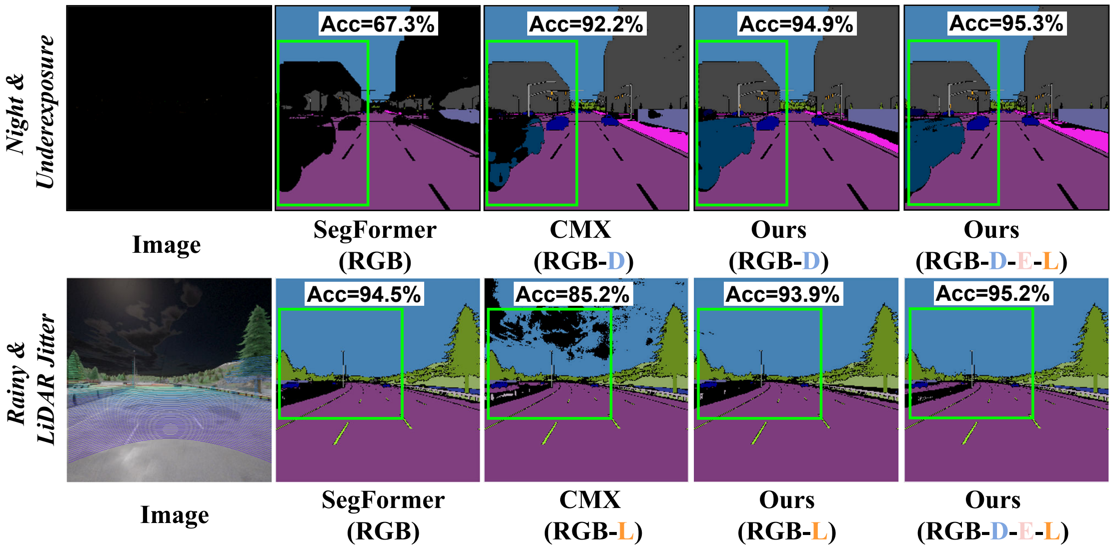


Figure 32: Visualization of segmentation results generated from DeLiVER dataset.

4.3 CHAPTER CONCLUSION

Robust scene understanding based on multimodal semantic segmentation represents a promising task, aiming to stabilize segmentation by fusing diverse modalities, including RGB images, depth maps, LiDAR point clouds, event-based data, *etc*. In the second research theme of ITS, we conduct an extensive study of multimodal semantic segmentation, concentrating on two primary aspects: *cross-modal* fusion and *arbitrary-modal* fusion. Within this chapter, we introduce two novel settings aimed at achieving robust scene understanding through panoramic semantic segmentation:

- **Cross-Modal Fusion for RGB-X Segmentation (CMX)**: A unified cross-modal fusion paradigm is a framework that can be used to extract complementary information from multiple data modalities. This framework is flexible enough to be applied to different combinations of data modalities, such as RGB-X (where X can represent any other modality, such as depth, thermal, LiDAR, and more).
- **Arbitrary-Modal Semantic Segmentation (AMSS)**: We take a step further in multimodal fusion by exploring arbitrary-modal semantic segmentation (AMSS), which can combine more data to combat against sensor failures and adverse cases. For example, AMSS can fuse a range of 2 to a total of 81 modalities, and it can be robust against adverse weather conditions such as fog and rain.

Here is a more detailed summary of the contributions of each section in this chapter:

Contribution 1: For the first time, we explore RGB-X semantic segmentation in five types of data combinations, including RGB-Depth, RGB-Thermal, RGB-Polarization, RGB-Event, and RGB-LiDAR. We propose a unified model *CMX* for cross-attention and it achieves state-of-the-art performances.

Contribution 2: We create a new benchmark DeLiVER for AMSS with four modalities, four adverse weather conditions, and five sensor failure modes. It includes six views with totally 47,310 frames, each of which contains semantic and instance segmentation labels.

Contribution 3: We present the Hub2Fuse paradigm with an asymmetric architecture to attain AMSS. Based on that, we propose a universal model *CMNeXt* that includes a Self-Query Hub (SQ-Hub) for selecting informative features and a Parallel Pooling Mixer (PPX) for harvesting discriminative cues.

Part III

MOBILITY ASSISTANCE SYSTEMS

5 TOWARDS NAVIGATIONAL SCENE UNDERSTANDING

In this chapter, we present the first research theme in MAS, *i.e.*, *navigational scene understanding*. MAS can help People with Visual Impairments (PVI) navigate by providing them with visual localization and semantic maps, which are key and fundamental components of navigation systems. For example, accessible and informative maps can help PVI to gain a clearer understanding of destinations before visiting, which can be helpful for planning trips and making informed decisions. However, robust indoor navigation and enhanced map accessibility for PVI are challenging due to the difficulty of perceiving low-texture indoor environments and constructing dense semantic maps. To tackle these issues, our contributions are three-fold: (1) We propose an efficient feature matching method (MatchFormer) for robust visual localization. This method is presented in Section 5.1, based on our work published in *ACCV 2022* [228]. (2) We create an end-to-end framework (Trans4Map) to build indoor Birds-Eye-View (BEV) semantic maps, which is presented in Section 5.2, based on our work published in *WACV 2023* [25]. (3) We pioneer a novel semantic mapping task, *i.e.*, 360° BEV mapping from a panorama, which is presented in Section 5.3, based on our work published in *WACV 2024* [204].

5.1 FEATURE MATCHING FOR VISUAL LOCALIZATION

This section is based on our work published in *ACCV 2022* [228]. The included methodological designs and technical contributions are collaborated in the context of a co-supervised master's thesis project.

5.1.1 *Interleaving Attention for Feature Matching*

Image feature matching is a key factor for visual localization and pose estimation, which is further essential for navigation assistance systems [26, 118, 148, 290] for People with Visual Impairments (PVI). Specifically, visual localization refers to the ability to determine the relative position of a camera or the user in a known environment, while pose estimation refers to the ability to determine the related orientation of the target user. To achieve these goals, image feature matching methods work by finding corresponding features between two or more images. These features can be points, lines, or other patterns. Once the corresponding features have been found, they can be used to estimate the transformation between the two images, which can then be

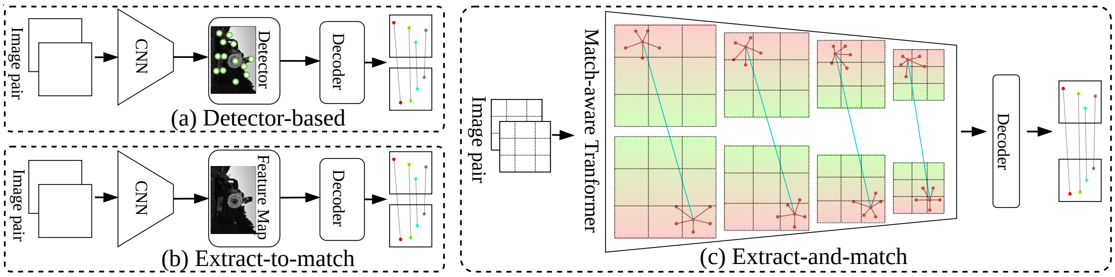


Figure 33: Feature matching pipelines. While (a) *detector-based* methods coupled with feature descriptors, (b) *extract-to-match* methods fail to make use of the matching capacity of the encoder. Self- and cross-attention are interleaved inside each stage of the match-aware transformer to perform a novel (c) *extract-and-match* pipeline.

used to determine the position and orientation of the camera or user. However, robust image feature matching for texture-less indoor scenarios are still under explored, but they have the potential to greatly improve the independence of PVI. In this section, we mainly explore to apply vision transformer [52] to build an efficient feature matching method for visual localization and pose estimation.

Specifically, for vision-based matching, classical *detector-based* methods (Figure 33a), coupled with hand-crafted local features [55, 168], are computationally intensive due to the high dimensionality of local features [173, 307]. Recent works [131, 161, 219] based on deep learning focus on learning detectors and local descriptors using Convolutional Neural Networks (CNNs). Some partial transformer-based methods [88, 194] only design an attention-based decoder and remain the *extract-to-match* pipeline (Figure 33b). For instance, while COTR [88] feeds CNN-extracted features into a transformer-based decoder, SuperGlue [174] and LoFTR [194] only apply attention modules atop the decoder. Overburdening the decoder, yet neglecting the matching capacity of the encoder, makes the whole model computationally inefficient.

Rethinking local feature matching, in reality, one can perform feature extraction and matching simultaneously by using a pure transformer. We propose an *extract-and-match* pipeline shown in Figure 33c. Compared to the *detector-based* methods and the *extract-to-match* pipeline, our new scheme is more in line with human intuition, which learns more respective features of image pairs while paying attention to their similarities [298]. Based on this, a novel method *MatchFormer* is proposed, which helps to achieve multi-wins in precision, efficiency, and robustness of feature matching.

More specifically, for improving computational efficiency and the robustness in matching low-texture scenes, we put forward *interleaving* self- and cross-attention in MatchFormer to build a matching-aware encoder. In this way, the local features of the image itself and the similarities of its paired images can be learned simultaneously, so called *extract-and-match*, which relieves the overweight decoder and makes the whole model efficient. The cross-attention arranged in earlier stages of the encoder robustifies feature matching, particularly, in low-texture indoor scenarios or with less training samples outdoors, which makes MatchFormer more suitable for real-world applications where large-scale data collection and annotation are infeasible.

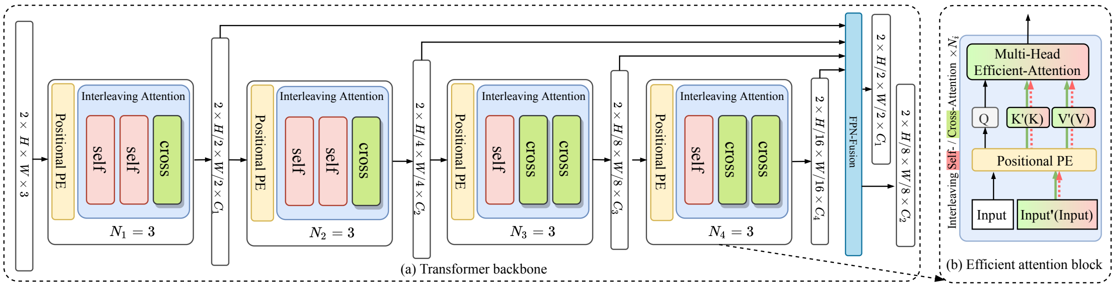

$2 \times H \times W \times 3$
Positional PE
Interleaving Attention
self
self
cross
$N_1 = 3$
$2 \times H/2 \times W/2 \times C_1$
Positional PE
Interleaving Attention
self
self
cross
$N_2 = 3$
$2 \times H/4 \times W/4 \times C_2$
Positional PE
Interleaving Attention
self
cross
cross
$N_3 = 3$
$2 \times H/8 \times W/8 \times C_3$
Positional PE
Interleaving Attention
self
cross
cross
$N_4 = 3$
$2 \times H/16 \times W/16 \times C_4$
FPN-Fusion
$2 \times H/2 \times W/2 \times C_1$
$2 \times H/8 \times W/8 \times C_2$
(a) Transformer backbone
Interleaving Self- / Cross-Attention $\times N_i$
Multi-Head Efficient-Attention
Q
K'(K)
V'(V)
Positional PE
Input
Input'(Input)
(b) Efficient attention block

Figure 34: MatchFormer architecture: (a) The transformer backbone generates high-resolution coarse features and low-resolution fine features; In (b), each attention block has interleaving-arranged self-attention (*w.r.t.* $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ and red arrows) within the $\mathtt{input}$, and cross-attention (*w.r.t.* $\boldsymbol{Q}$, $\boldsymbol{K'}$, $\boldsymbol{V'}$ and alternative green arrows) cross images ($\mathtt{input}$ and $\mathtt{input}'$). Multi-head efficient-attention reduces the computation; The positional Patch Embedding (PE) completes the patch embedding and the position encoding.

5.1.2 *MatchFormer Model*

As illustrated in Figure 34, MatchFormer employs a hierarchical transformer, which comprises four stages to generate high-resolution coarse and low-resolution fine features for local feature matching. In four stages, the self- and cross-attention modules are arranged in an *interleaving* strategy. Each stage consists of two components: one *positional patch embedding (PosPE)* module, and a set of efficient attention modules. Then, the multi-scale features are fused by an FPN-like decoder, which are passed to perform the coarse-to-fine matching, following [194].

Extract-and-Match Pipeline. Unlike the *extract-to-match* LoFTR using attention on a single-scale feature map and only after feature extraction, we combine self- and cross-attention inside the encoder and apply on multiple feature scales (Figure 33). The combination of two types of attention modules enables the model to extract non-local features via self-attention and explore their similarities via cross-attention simultaneously, so called the *extract-and-match* scheme.

Interleaving Self-/Cross-Attention. As shown in Figure 34a, the combination of self- and cross-attention modules are set at each stage in an *interleaving* strategy. Each block in Figure 34b contains N attention modules, where each attention module is represented as self-attention or alternative cross-attention according to the input image pair. For self-attention, $\boldsymbol{Q}$ and $(\boldsymbol{K}, \boldsymbol{V})$ come from the same *input*, so the self-attention is responsible for feature extraction of the image itself. For cross-attention, $(\boldsymbol{K'}, \boldsymbol{V'})$ are from another $\mathtt{input}'$ of the image pair. Thus, the cross-attention learns the similarity of the image pair, resulting in a match-aware transformer-based encoder.

MatchFormer Variants. MatchFormer is available with its *lite* and *large* versions, as presented in Table 16. We set MatchFormer-lite 4-stage features in the respective resolution of $\frac{1}{r_i} \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of the input. To promote context learning for matching, feature embeddings with higher channel numbers are beneficial, which are set as $C_i \in \{128, 192, 256, 512\}$ for four stages. In the MatchFormer-large models, higher resolution feature maps facilitate accurate dense matching. Hence, the $\frac{1}{r_i}$ and C_i are set as $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ and $\{128, 192, 256, 512\}$ for the large MatchFormer.

Stage	MatchFormer-lite		MatchFormer-large		N_i
$\boldsymbol{F}_1$	H/4×W/4	K=7, S=4, P=3, E=4	H/2×W/2	K=7, S=2, P=3, E=4	
	C_1=128	**LA:** A=8 ; **SEA:** A=1, R=4	C_1=128	**LA:** A=8 ; **SEA:** A=1, R=4	×3
$\boldsymbol{F}_2$	H/8×W/8	K=3, S=2, P=1, E=4	H/4×W/4	K=3, S=2, P=1, E=4	
	C_2=192	**LA:** A=8 ; **SEA:** A=2, R=2	C_2=192	**LA:** A=8 ; **SEA:** A=2, R=2	×3
$\boldsymbol{F}_3$	H/16×W/16	K=3, S=2, P=1, E=4	H/8×W/8	K=3, S=2, P=1, E=4	
	C_3=256	**LA:** A=8 ; **SEA:** A=4, R=2	C_3=256	**LA:** A=8 ; **SEA:** A=4, R=2	×3
$\boldsymbol{F}_4$	H/32×W/32	K=3, S=2, P=1, E=4	H/16×W/16	K=3, S=2, P=1, E=4	
	C_4=512	**LA:** A=8 ; **SEA:** A=8, R=1	C_4=512	**LA:** A=8 ; **SEA:** A=8, R=1	×3
Output	Coarse: H/4×W/4, 128 Fine: H/8×W/8, 192		Coarse: H/2×W/2, 128 Fine: H/8×W/8, 256		

Table 16: MatchFormer-lite and -large with Linear Attention (LA) and Spatial Efficient Attention (SEA). C: the channel number of feature $\boldsymbol{F}$; K, S and P: the patch size, stride, and padding size of PosPE; E: the expansion ratio of MLP in an attention block; A: the head number of attention; R: the down-scale ratio of SEA.

Attention Module Variants. To fully explore the proposed *extract-and-match* scheme, each of the two MatchFormer variants has two attention variants. Here, we mainly investigate Linear Attention (LA) and Spatial Efficient Attention (SEA). Thus, there are four versions of MatchFormer as presented in Table 16.

5.1.3 *Experiments and Analysis*

Metrics. Following [174], we provide the area under the cumulative curve (AUC) of the pose error at three different thresholds (5°, 10°, 20°). The camera pose is recovered by using RANSAC. We report the matching precision (P), the probability of a true match if its epipolar is smaller than 5×10^{-4} in indoor cases and 1×10^{-4} in outdoor cases [174].

5.1.3.1 *Indoor Pose Estimation*

Indoor pose estimation is a key components of indoor navigation systems. MatchFormer with interleaved self- and cross-attention modules functions well, delivering a promising navigational scene understanding approach to mobility assistance systems. **Quantitative Results.** As shown in Table 17, MatchFormer demonstrates exceptional performance on the low-texture indoor pose estimation task. The matching precision (P) of MatchFormer-large-SEA reaches the state-of-the-art level of 89.5%. Benefiting from the *extract-and-match* strategy, MatchFormer-large-SEA can bring +5.1% improvement over the *detector-based* SuperGlue, +1.6% over the *extract-to-match* LoFTR. Pose estimation area under the curve (AUC) of MatchFormer is superior to *detector-based* SuperGlue. Compared to LoFTR, MatchFormer provides a more pronounced pose estimation AUC by boosting (+2.25%, +3.1%, +3.79%) at three thresholds of (5°, 10°, 20°). The LoFTR model is recently adapted by a complex decoder with QuadTree Attention [202]. However, MatchFormer maintains its lead (+0.41%, +0.70%, +1.11%) with the *extract-and-match* strategy. Compared to LoFTR, our lightweight MatchFormer-lite-SEA has only 45% GFLOPs, yet achieves a +1.3% precision gain and a 41% running speed boost. Comparing SEA and LA, we found that the spatial scaling operation in SEA has benefits in handling low-texture features in indoor scenes.

Method	Pose estimation AUC (%)			P
	@5°	@10°	@20°	
ORB [168]+GMS [12] CVPR'17	5.21	13.65	25.36	72.0
D2-Net [55]+NN CVPR'19	5.25	14.53	27.96	46.7
ContextDesc [131]+RT [128] CVPR'19	6.64	15.01	25.75	51.2
SP [50]+NN CVPRW'18	9.43	21.53	36.40	50.4
SP [50]+PointCN [254] CVPR'18	11.40	25.47	41.41	71.8
SP [50]+OANet [277] ICCV'19	11.76	26.90	43.85	74.0
SP [50]+SuperGlue [174] CVPR'20	16.16	33.81	51.84	84.4
LoFTR [194] CVPR'21	22.06	40.80	57.62	87.9
LoFTR [194]+QuadTree [202] ICLR'22	23.90	43.20	60.30	89.3
MatchFormer-lite-LA	20.42	39.23	56.82	87.7
MatchFormer-lite-SEA	22.89	42.68	60.66	89.2
MatchFormer-large-LA	24.27	43.48	60.55	89.2
MatchFormer-large-SEA	**24.31**	**43.90**	**61.41**	**89.5**

Table 17: Indoor pose estimation on ScanNet. The AUC of three different thresholds and the average matching precision (P) are evaluated.

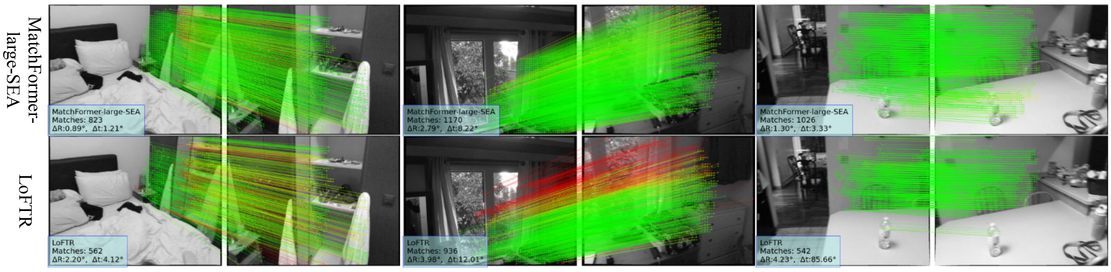


Figure 35: Qualitative visualization of MatchFormer and LoFTR [194]. MatchFormer achieves higher matching numbers and more correct matches in low-texture scenes.

Qualitative Results. The indoor matching results are in Figure 35. In challenging feature-sparse indoor scenes, it can reliably capture global information to assure more matches and high accuracy. Thus, the pose solved by matching prediction has a lower maximum angle error (ΔR) and translation error (Δt). The result confirms that applying cross-attention modules earlier for learning feature similarity robustifies low-texture indoor matching, which is in line with our *extract-and-match* pipeline.

5.1.3.2 *Outdoor Pose Estimation*

Outdoor pose estimation presents unique challenges compared to indoors. In particular, outdoor scenes have greater variations in lighting and occlusion. Still, Matchformer achieves outstanding performance in outdoor scenes.

Quantitative Results. As shown in Table 18, MatchFormer noticeably surpasses the *detector-based* SuperGlue and DRC-Net, as well as the *extract-to-match* LoFTR. Our MatchFormer-lite-LA model achieves a higher matching precision (P) with 97.55%, despite being much lighter. Note that MatchFormer-large-SEA using the partially optimized SEA will raise an out-of-memory issue. Here, we recommend to use the memory-efficient LA in the high-resolution outdoor scenes. Our MatchFormer-large-LA model achieves consistent state-of-the-art performances on both metrics of AUC and P.

Robustness and Resource-Efficiency. It is important to evaluate the model robustness when less training data and fewer training resources are available in practical applications. Thus, we train MatchFormer-large-LA and LoFTR (marked with † in Table 18) using different percentages of datasets and on fewer resources (8 GPUs). First,

Method	Data percent	Pose estimation AUC (%)			P
		@5°	@10°	@20°	
SP [50]+SuperGlue [174] CVPR'20	100%	42.18	61.16	75.95	–
DRC-Net [107] NeurIPS'20	100%	27.01	42.96	58.31	–
LoFTR [194] CVPR'21	100%	52.80	69.19	81.18	94.80
MatchFormer-lite-LA	100%	48.74	65.83	78.81	97.55
MatchFormer-lite-SEA	100%	48.97	66.12	79.07	97.52
MatchFormer-large-LA	100%	**52.91** (+0.11)	**69.74** (+0.55)	**82.00** (+0.82)	**97.56** (+2.76)
Robustness with less training data and fewer GPU resources:					
LoFTR†	10%	38.81	54.53	67.04	83.64
MatchFormer†	10%	42.92 (+4.11)	58.33 (+3.80)	70.34 (+3.30)	85.08 (+1.44)
LoFTR†	30%	47.38	64.77	77.68	91.94
MatchFormer†	30%	49.53 (+2.15)	66.74 (+1.97)	79.43 (+1.75)	94.28 (+2.34)
LoFTR†	50%	48.68	65.49	77.62	92.54
MatchFormer†	50%	50.13 (+1.45)	66.71 (+1.22)	79.01 (+1.39)	94.89 (+2.35)
LoFTR†	70%	49.08	66.03	78.72	93.86
MatchFormer†	70%	51.22 (+2.14)	67.44 (+1.41)	79.73 (+1.01)	95.75 (+1.89)
LoFTR†	100%	50.85	67.56	79.96	95.18
MatchFormer†	100%	**53.28** (+2.43)	**69.74** (+2.18)	**81.83** (+1.87)	**96.59** (+1.41)

Table 18: Outdoor pose estimation on MegaDepth. † represents training on different percentages of datasets, which requires 8 GPUs for training.

compared to LoFTR†, MatchFormer† obtains consistent improvements on different scales, *i.e.*, the first {10,30,50,70,100} percentages of the original dataset. It proves that MatchFormer has more promise in data-hungry real-world applications. Second, training with the same 100% data on different GPU resources, LoFTR† has (-1.95%, -1.63%, -1.22%) performance drops at three AUC thresholds of (5°, 10°, 20°) when using 8 GPUs instead of 64 GPUs. In contrast, MatchFormer maintains the stable and surprising accuracy, which shows that our method is more resource-friendly and easier to reproduce.

5.1.3.3 *Visual Localization on InLoc*

As shown in Table 19, on the InLoc benchmark for visual localization, MatchFormer reaches a level comparable to the current state of art methods SuperGlue and LoFTR. Interleaving attention in the MatchFormer backbone enables robust local feature matching in indoor scenes with large low-texture areas and repetitive structures.

Method	Localized Queries (%, 0.25m/0.5m/1.0m, 10°)	
	DUC1	DUC2
SP [50] + NN CVPRW'18	40.4 / 58.1 / 69.7	42.0 / 58.8 / 69.5
D2Net [55] + NN CVPR'19	38.4 / 56.1 / 71.2	37.4 / 55.0 / 64.9
R2D2 [161] + NN NeurIPS'19	36.4 / 57.6 / 74.2	45.0 / 60.3 / 67.9
SP [50] + SuperGlue [174] CVPR'20	49.0 / 68.7 / 80.8	53.4 / **77.1** / 82.4
SP [50] + CAPS [219] + NN ECCV'20	40.9 / 60.6 / 72.7	43.5 / 58.8 / 68.7
SP [50] + ClusterGNN [183] CVPR'22	47.5 / 69.7 / 79.8	53.4 / **77.1** / 84.7
ASLFeat [132] + SuperGlue [174] CVPR'20	51.5 / 66.7 / 75.8	53.4 / 76.3 / 84.0
ASLFeat [132] + ClusterGNN [183] CVPR'22	**52.5** / 68.7 / 76.8	55.0 / 76.0 / 82.4
SIFT + CAPS [219] + NN ECCV'20	38.4 / 56.6 / 70.7	35.1 / 48.9 / 58.8
SparseNCNet [162] ECCV'20	41.9 / 62.1 / 72.7	35.1 / 48.1 / 55.0
Patch2Pix [304] CVPR'21	44.4 / 66.7 / 78.3	49.6 / 64.9 / 72.5
LoFTR-OT [194] CVPR'21	47.5 / 72.2 / 84.8	54.2 / 74.8 / **85.5**
MatchFormer	46.5 / **73.2** / **85.9**	**55.7** / 71.8 / 81.7

Table 19: Visual localization on InLoc. We report the percentage of correctly localized queries under specific error thresholds, following the HLoc [173] pipeline.

5.2 BIRD'S-EYE-VIEW SEMANTIC MAPPING

This section is based on our work published in *WACV 2023* [25]. The included methodological designs and technical contributions are collaborated in the context of a co-supervised master's thesis project.

5.2.1 *Top-down Semantic Mapping Pipelines*

Semantic mapping is still difficult for an artificial intelligent mobile agent, particularly when exploring an unfamiliar environment. Besides, semantic mapping can generate accessible maps to help PVI to better understand their destination before visiting. Through beforehand information and planning obtained in advance, independent on-site navigation of PVI can be further improved. In this work, we focus on image-based semantic mapping task, by predicting allocentric semantic segmentation from given egocentric images. As shown in Figure 36, given a trajectory in the scene, which is composed of a batch of first-view RGB images and the corresponding known camera pose, the mobile agent performs three steps: (1) extracting rich and compact contextual features; (2) projecting and updating the egocentric features in the allocentric memory as spatial-semantic representation; (3) decoding final top-view semantic mapping. The image-based egocentric-to-allocentric mapping pipeline is more in line with human intuition, and is able to perform mapping in an efficient way, avoiding the need for a time-consuming reconstruction phase [68].

Image-based semantic mapping methods can be divided into 4 pipelines, as summarized in Figure 37. The project-then-segment pipeline (Figure 37a) projects N observations into Bird's Eye View (BEV), that hinders the small object segmentation due to the lack of fine visual information. The segment-then-project pipeline (Figure 37b) depends heavily on the front-view segmentation performance and may accumulate errors. The offline project-then-segment pipeline (Figure 37c) requires large-scale lo-

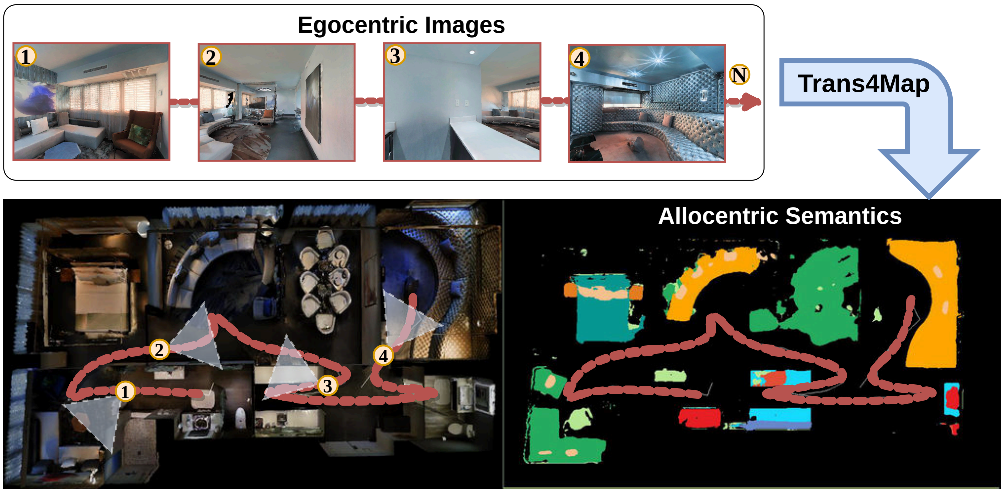
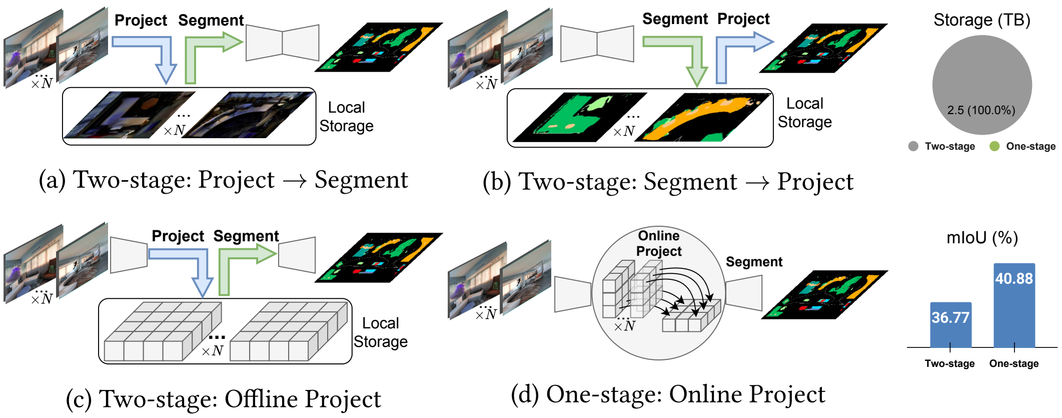


Figure 36: The egocentric-to-allocentric semantic mapping. Given a front-view image sequence of length N observed along a trajectory (the **red** dash line), Trans4Map performs the online extract-project-segment pipeline, yielding an allocentric semantic map in bird's eye view.



(a) Two-stage: Project → Segment

(b) Two-stage: Segment → Project

(c) Two-stage: Offline Project

(d) One-stage: Online Project

Figure 37: Semantic mapping pipelines. The two-stage pipelines in (a)(b)(c) differ from the projection locations, *i.e.*, early, late, and intermediate offline projection. The one-stage pipeline in (d) reduces 2.5TB (100%) of storage to 0TB by using online projection and has a higher mIoU.

cal storage to save the feature map given by the pre-trained encoder of the first stage. It further demands huge GPU memory to reload offline features for the second stage training. Unlike two-stage pipelines above, our proposed online project-then-segment pipeline (Figure 37d) performs online implicit projection and enables end-to-end and resource-friendly BEV semantic mapping. The one-stage pipeline is crucial, because it fits resource-limited platforms, *e.g.*, robots and assistive systems. Further, it helps mobile agents to quickly construct maps and get familiar with the unknown space.

However, a lightweight but effective backbone that requires few resources is the decisive factor to achieve successful one-stage semantic mapping. The vision transformer architecture [212] is able to capture long-distance contextual dependencies, forming a non-local representation. This mechanism naturally fits the semantic mapping task, since the mapping process demands a holistic understanding of scenes. This assumption leads us to revisit the top-down semantic mapping with a transformer-based model and put forward a novel end-to-end one-stage *Trans4Map* framework. It delivers two primary benefits: (1) the long-range feature modeling ability is advantageous to obtain a more comprehensive spatial representation during the egocentric observation process; (2) the efficient and lightweight model structure enables the one-stage end-to-end mapping pipeline. Besides, unlike the previous method [22] using a single GRU cell to reload the offline features, we propose a novel *Bidirectional Allocentric Memory (BAM)* to combine features from both directions, which can avoid the occluded objects to be classified as other category, *e.g.*, *chairs* under *tables*. Further, our BAM implicitly performs the efficient online projection, as another key point in implementing the one-stage mapping pipeline (Figure 37d).

5.2.2 *Trans4Map Model*

As shown in Figure 38, our end-to-end Trans4Map framework includes three steps: (1) the incoming N egocentric images are fed into the transformer-based backbone (in Section 5.2.2.2), which extracts contextual feature and long-range dependency; (2) the Bidirectional Allocentric Memory (BAM) module (in Section 5.2.2.3) projects the

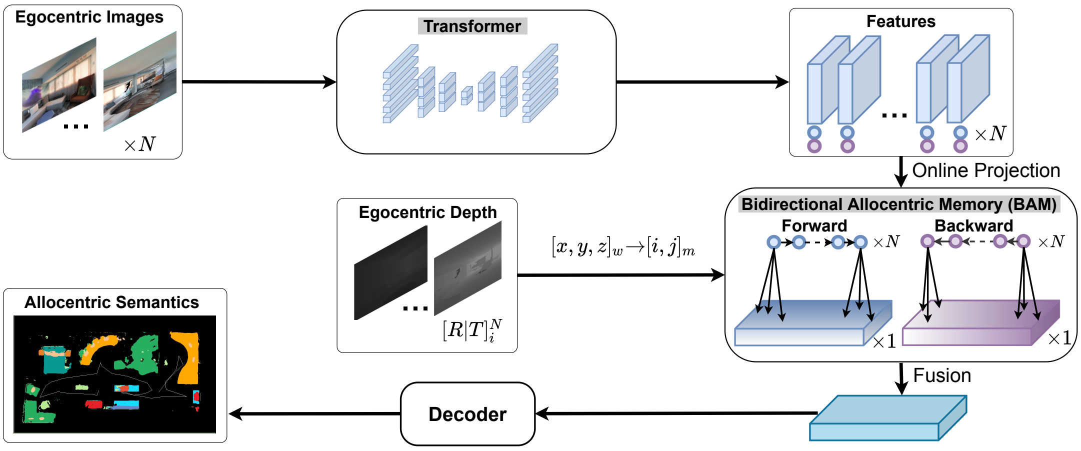


Figure 38: The overview of the end-to-end Trans4Map framework. There are a transformer-based encoder for extracting the egocentric features from the RGB images, a Bidirectional Allocentric Memory (BAM) to project and accumulate the extracted feature sequence to the allocentric feature map via the known depth and pose information, and a CNN-based decoder for parsing the accumulated feature and predicting the allocentric semantics.

extracted feature via the depth-based transformation index; (3) the lightweight CNN-based decoder parses the projected feature and predicts the allocentric semantics.

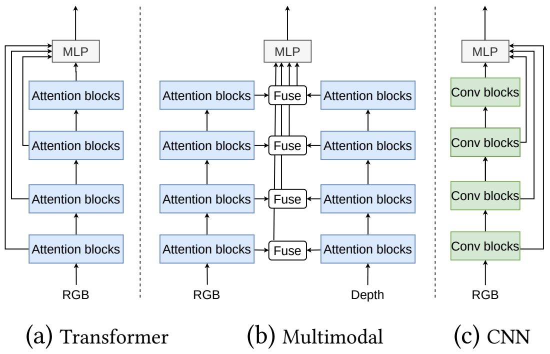
5.2.2.1 *One-stage Pipeline*

Unlike multi-stage methods [22], our framework operates in an one-stage end-to-end manner, benefiting from three designs: (1) a transformer-based backbone is leveraged to capture holistic features and long-range dependencies, instead of narrow-receptive-field CNN-based backbones; (2) a single branch structure for extracting RGB features makes the whole model more lightweight than the dual-branch one; (3) an online training pipeline from egocentric images to allocentric semantics is constructed, avoiding using the time-consuming two-stage process and feature maps storage.

5.2.2.2 *Transformer Backbone*

To fully investigate the proposed Trans4Map framework, we explore different model architectures and learning modalities, as shown in Figure 39. The architectures are constructed by four stages, and each stage includes a series of convolutional blocks (Figure 39c) or self-attention blocks (Figure 39a). Considering that cross-modality complementary features are informative for predicting semantics [82, 87, 278], we leverage RGB-Depth inputs and a multimodal architecture (Figure 39b) is reformed by using efficient self-attention blocks.

For brevity, we describe the operation of the single-modal process, while the bimodal process involves an additive fusion at each stage, in which the fusion block obtains the extracted contextual features and geometry features and then fuses them per pixel with the same dimension. Given a batch of RGB images of size $N \times H \times W \times 3$, the divided patches are passed through the four-stage transformer blocks, to obtain the hierarchical feature representation with downsampling rates of $\{\frac{1}{r_1}, \frac{1}{r_2}, \frac{1}{r_3}, \frac{1}{r_4}\}$ and increasing channels of $\{C_1, C_2, C_3, C_4\}$. Then, the multi-scale features are concatenated by an



(a) Transformer (b) Multimodal (c) CNN

Figure 39: Semantic mapping architectures.

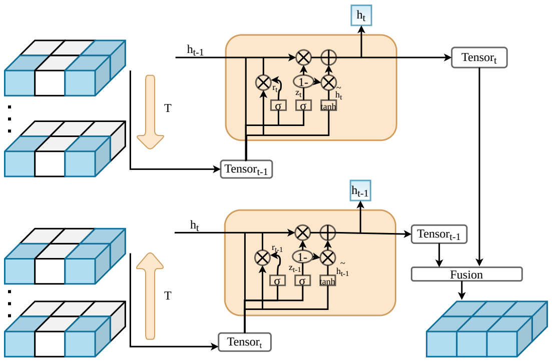


Figure 40: Bidirectional Allocentric Memory.

MLP layer and followed by a convolution layer with 64 channels. So the hierarchical features are fused into an egocentric feature of size $N \times \frac{H}{r_1} \times \frac{W}{r_1} \times 64$. To study different semantic mapping architectures, the multi-scale features in this work are extracted with $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ downsampling rates and $\{64, 128, 320, 512\}$ channels.

To compare CNN- and transformer-based models, the RedNet backbone [87] used in SMNet and the ConvNeXt backbone are selected to form CNN-based mapping models, while transformer-based models include FAN [300], Swin [125], and SegFormer [239] backbones. Based on our experiments, we adopt SegFormer [239] as the default backbone of our visual encoder, as its simple and lightweight design can generate features ranging from high-resolution fine features to low-resolution coarse features.

5.2.2.3 *Bidirectional Allocentric Memory*

After acquiring the egocentric features through the aforementioned transformer backbone, the projective index is needed to project representative contextual features into an allocentric memory map. In the Habitat simulator [175], we can directly obtain the state of the moving agent and then calculate the camera pose using relative orientation and position. In order to perform the online projection, we need to derive the 3D position of each pixel in the egocentric image, as presented as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_c = K^{-1} \begin{bmatrix} u \\ v \\ d_{u,v} \end{bmatrix}_i , \quad \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_w = R^{-1} \begin{bmatrix} x \\ y \\ z \end{bmatrix}_c - \overrightarrow{t} . \tag{31}$$

K is the camera intrinsic parameter, $[R|\overrightarrow{t}]$ are the rotation matrix and the translation matrix, respectively. First, in Eq. (31), using the camera model and the depth of each pixel $d_{u,v}$, the pixel coordinate (u, v) in the image coordinate system can be converted into the camera coordinate system. Then, the camera coordinates denoted as (x, y, z) of each point are converted to world coordinates denoted as (X, Y, Z) using the rotation matrix and translation matrix. Each pixel in the allocentric memory map represents a 2cm×2cm cell in the scene of Matterport3D dataset [23], so the projective index $(i, j)_m$ can be calculated by dividing X and Z. Finally, we project egocentric features of N batch size onto the allocentric memory map using the calculated projective index.

To enhance the long-range content dependency and aggregate the incoming information completely, we propose *Bidirectional Allocentric Memory* (*BAM*), in which we use Bi-directional GRU (BiGRU) to update incoming observations from two directions. As shown in Figure 40, the upper GRU processes the allocentric memory tensor in a forward direction ($M^{t-1} \rightarrow M^{t}$) and the lower GRU in a backward direction ($M^{t} \rightarrow M^{t-1}$). A simple yet effective convolutional layer is applied to fuse two projected features. The computation of updated spatial memory tensor is formulated as:

$$M_{i,j}^{t} = \mathrm{GRU}(F_{i,j}^{t}, M_{i,j}^{t-1}); \tag{32}$$

$$M_{i,j}^{t-1} = \mathrm{GRU}(F_{i,j}^{t-1}, M_{i,j}^{t}); \tag{33}$$

$$T = \mathrm{Conv}(M_{i,j}^{t}, M_{i,j}^{t-1}). \tag{34}$$

$M_{i,j}^{t}$ and $M_{i,j}^{t-1}$ are the current time step spatial memory and previous time step spatial memory, respectively. The fused spatial memory tensors T are accessible to the decoding step for the final semantic top-down map prediction. Thanks to the bidirectional parsing process, BAM is able to accumulate the observations per each time step in both direction in parallel, thus, it can better avoid occluded objects being wrongly-classified. With BAM, Trans4Map can produce a more meaningful allocentric representation which combines bidirectional projected features.

5.2.3 *Experiments and Analysis*

Matterport3D. The results on Matterport3D are in Table 20. Following the segment-project paradigm, the result obtained by using the label data is the upper bound performance. As in Table 20, the segment-project baseline performs much better than the project-segment one, since part of information will be lost in the process of converting an egocentric image into a top-down view. The semantic SLAM in [68] also uses the segment-project method but achieves worse performance than the image-based segment-project baseline. SMNet [22] follows the offline project-segment paradigm and adds a spatial memory update module. Here, we reproduce the experiment using the released code under the same condition and obtain the results with a mIoU score of 36.16% and a mBF1 value of 35.95%. Compared to SMNet, our Trans4Map model achieves significant improvements in terms of mIoU (40.02%) and mBF1 (41.11%) on the Matterport3D dataset, which proves the effectiveness of our framework.

In Figure 41, we present a comparison with advanced architectures [125, 126, 238, 300]. We found that simply applying transformer-based backbones does not guarantee improvement. Thanks to the proposed method, our Trans4Map models have much

Method	**Acc**	**mRecall**	**mPrecision**	**mIoU**	**mBF1**
Seg. GT → Proj.	89.49	73.73	74.58	59.73	54.05
Two-stage Proj. → Seg.	83.18	27.32	35.30	19.96	17.33
Two-stage Seg. → Proj.	88.06	40.53	**58.92**	32.76	33.21
Two-stage Semantic SLAM	85.17	37.51	51.54	28.11	31.05
Two-stage SMNet	88.14	47.49	58.27	36.77	37.02
Two-stage SMNet †	**89.14**	46.34	56.98	36.16	35.95
One-stage Trans4Map	89.02	**54.50**	56.20	**40.02**	**41.11**

Table 20: Allocentric semantic mapping results on Matterport3D. † is our reproduction.

Method	**Acc**	**mRecall**	**mPrecision**	**mIoU**	**mBF1**
Seg. GT → Proj.	96.83	83.84	94.05	79.76	86.89
Two-stage Seg. → Proj.	88.61	48.11	65.20	40.77	45.86
Two-stage Semantic SLAM	88.30	45.80	62.41	37.99	46.71
Two-stage SMNet	89.26	53.37	64.81	43.12	45.18
Two-stage SMNet †	**87.69**	58.88	34.85	27.68	42.67
One-stage Trans4Map	86.19	**65.27**	**34.91**	**29.15**	**48.66**

Table 21: Allocentric semantic mapping results on Replica. † is our reproduction.

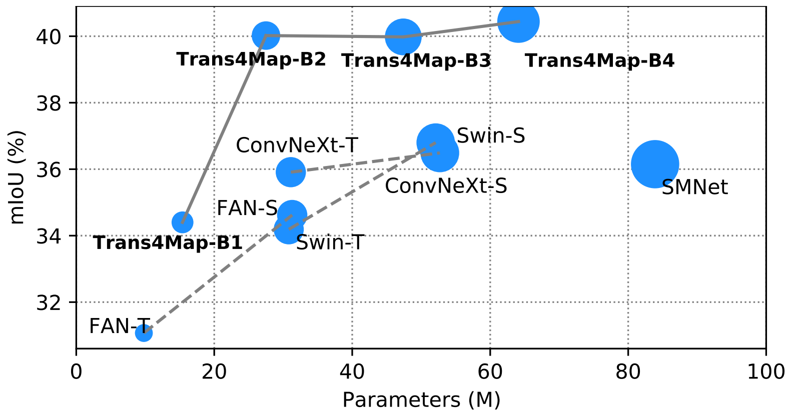


Figure 41: Semantic mapping scores (mIoU) using CNN and transformer backbones with different #parameters (M). Trans4Map models achieve better results yet with fewer parameters.

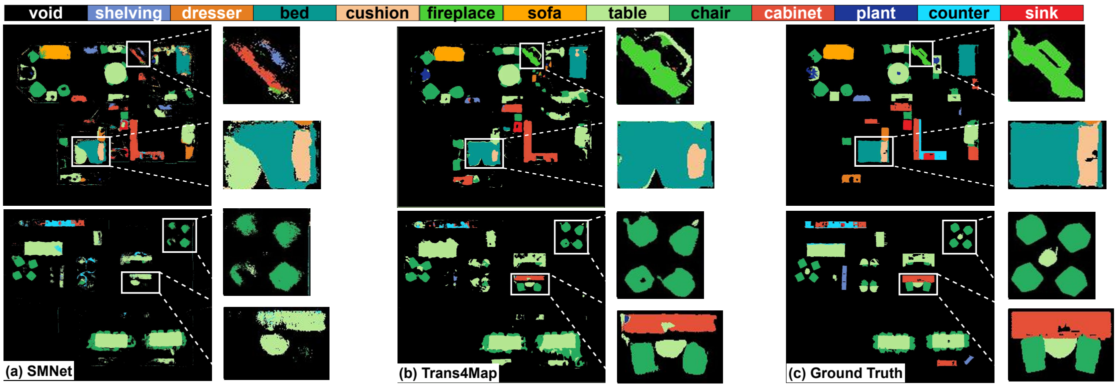


Figure 42: Allocentric semantic mapping visualizations. There are two indoor scenes from the Matterport3D test set. From left to right are the predicted results of SMNet, the results of our Trans4Map and the ground truth. Zoom in for better view.

fewer parameters, yet achieve better scores. The B2 version reduces 67.2% parameters over SMNet [22] and achieves >40% mIoU on Matterport3D [23].

Replica. The results on Replica are in Table 21. Note that the last two rows are evaluated on the partially available Replica [190] dataset, while the others have all data as [22]. All models are trained on the Matterport3D dataset and tested on the Replica dataset. The trajectories and labels of the Replica dataset are partially available at the moment, thus, the results are tested on the constrained data of the Replica dataset. Nonetheless, under the same condition with the same label data, our Trans4Map outperforms the baseline SMNet with a 1.47% mIoU and a 5.99% mBF1 improvements, respectively. The results indicate that our Trans4Map framework achieves consistent improvements across different datasets.

Semantic map visualizations. In Figure 42, we visualize the semantic map results from the test set of the Matterport3D dataset. Thanks to the extracted non-local features and long-distance dependencies, Trans4Map has better segmentation results. In the first scene in Figure 42, Trans4Map is better at segmenting the *bed*. Further, Trans4Map is able to successfully classify the *fireplace*, while the baseline model fails and predicts it as a *cabinet*. In the second scene in Figure 42, Trans4Map delivers semantic mapping accurately, such as on *cabinet* and *chair* categories, while SMNet misclassifies them as *tables*. Besides, SMNet yields incomplete *chair* segmentation results.

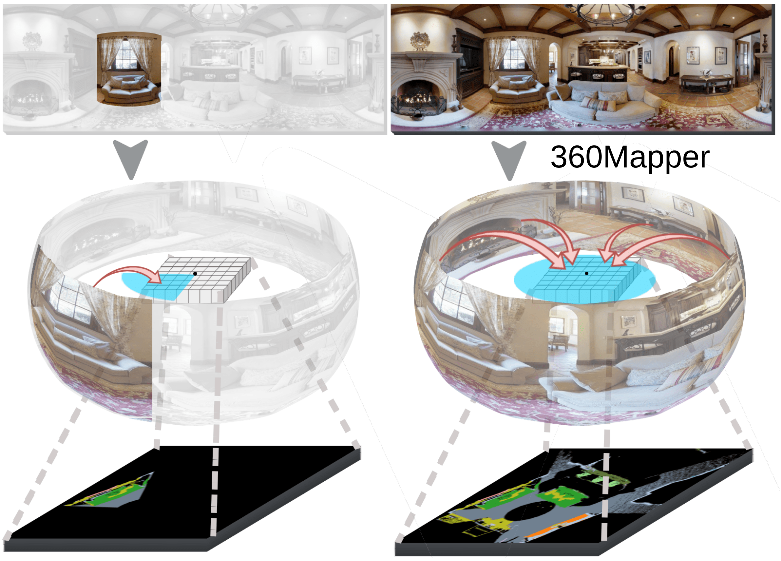
5.3 PANORAMIC SEMANTIC MAPPING

This section is based on our work published in *WACV 2024* [204]. The included methodological designs and technical contributions are collaborated in the context of a co-supervised master's thesis project.

5.3.1 *Semantic Mapping Using a Single Image*

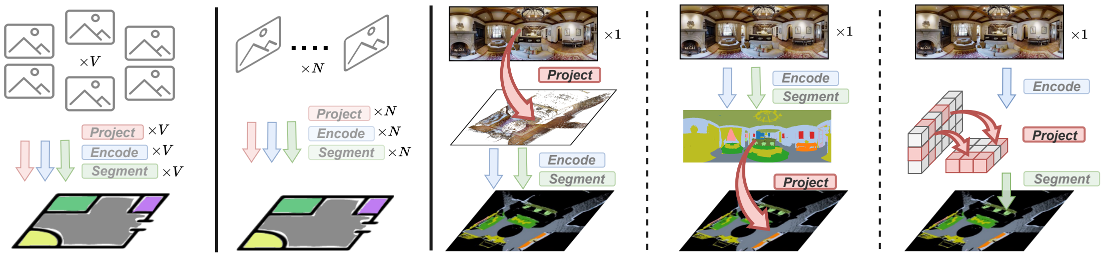
For semantically mapping front-view indoor scenes, sequence-based methods [22, 25] were proposed, which have to process whole videos and entail a moving camera. As shown in Figure 43a, (1) these methods rely on computationally expensive processing of entire sequences of video-frames due to the narrow field of view of the pinhole camera, and (2) they are constrained to explore indoor mapping on synthetic simulators [175, 234], due to the lack of real indoor datasets. These drawbacks limit their applicability to real-world indoor semantic mapping.

To solve these limitations, we introduce *360BEV* to achieve 360° semantic mapping for indoor BEV, which is illustrated in Figure 43b. Our considerations are two-fold: (1) To unleash the potential of indoor semantic mapping in real-world scenarios, real indoor databases with BEV semantic labels are crucial; (2) To reduce the computational complexity of narrow-FoV sequence methods [22] ($\geqslant$20 video-frames to process) or the complexity of multi-camera setups [111] ($\geqslant$6 camera views needed), we leverage a single-frame 360° image with depth information and thus bypass multi-sensor calibration, synchronization, and data fusion procedures. By decoupling the computationally expensive processing of sequences or multiple views, our direct 360BEV semantic mapping is more streamlined for generating indoor semantic maps.



(a) Narrow BEV (b) 360BEV

Figure 43: Semantic mapping from egocentric images to allocentric BEV semantics. While (a) the narrow-BEV method has limited perception and map range, (b) 360BEV has an omnidirectional field of view, yielding a more complete BEV map by using our 360Mapper model.



(a) Multi-view (b) Sequence-based (c) Early-projection (d) Late-projection (e) Intermediate-projection

Figure 44: Paradigms of semantic mapping. While the narrow-FoV (a) multi-view and (b) sequence-based methods rely on V⩾6 and N⩾20 views, the 360°-BEV (c) Early-, (d) Late-, and (e) Intermediate-projection methods use a single panorama.

5.3.2 *Benchmarking Panorama Semantic Mapping*

To investigate the 360BEV task, we analyze potential panoramic projection paradigms in Section 5.3.2.1. The data generation is detailed in Section 5.3.2.2.

5.3.2.1 *The 360 Projection Paradigms*

As shown in Figure 44, unlike multi-view methods (V⩾6 in Figure 44a) and sequence-based methods (N⩾20 in Figure 44b), 360BEV uses a single image with depth. We investigate three projection paradigms, *i.e.*, *how to process data from front-view panoramas to bird's-eye-view semantics*, which are:

1. *Early projection: **Project**→Encode→Segment* in Figure 44c, *i.e.*, view projection is first done in RGB images. This way of processing might harm the original visual information and the spatial relationship of indoor objects, leading to lower performance of semantic mapping.
2. *Late projection: Encode→Segment→**Project*** in Figure 44d, *i.e.*, feature extraction and segmentation are first done and view projection is executed at the end. The front-view segmentation errors caused by distortion and deformation of panoramas accumulate and affect the completeness of object masks in the BEV map.
3. *Intermediate projection: Encode→**Project**→Segment* in Figure 44e, *i.e.*, feature extraction and view projection are first done in order and segmentation is executed at the end. In this manner, the encoded feature maintains dense and representative information, which is crucial for view projection. Besides, the projected features are further parsed by the subsequent BEV decoder.

Based on these properties, we mainly explore 360BEV with intermediate projections, in which we identify the following challenges: In the feature extraction stage, spatial distortions and object deformations severely hinder the encoder from extracting representative features from the front-view panoramic image. For the intermediate feature projection, only depth information is utilized to consistent view transformation of high-dimensional features. In addition, many large objects in the front view (*e.g.*, *walls*) are projected to thin objects in the top-down view, which greatly impedes capturing wide-range features during projection.

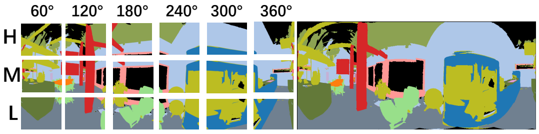


Figure 45: 360FV semantics generation from 18 narrow views to a panoramic view on the 360FV-Matterport dataset. H, M, L represent high, medium, and low positions, respectively.

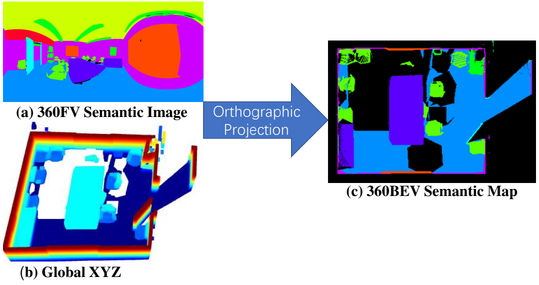


Figure 46: 360BEV semantics generation by orthographic projection, from (a) the front-view semantic image and (b) the global XYZ image, to (c) the 360BEV semantic map.

5.3.2.2 *Data Generation*

360FV-Matterport. The original Matterport3D [23] was collected via narrow-FoV cameras. As shown in Figure 45, we convert the 18 narrow-view images and annotations into the 360° format by using rotation-translation matrices.

360BEV-Stanford. The Stanford2D3D dataset [5] has front-view labels but not BEV labels. As presented in Figure 46, we utilize the global XYZ image to generate the corresponding BEV semantic map by applying orthographic projection.

360BEV-Matterport. For Matterport dataset, we generate global XYZ images by using the depth ground truth. First, a panoramic image can be processed as a sphere with rays shooting from the center of the sphere, where the camera is located.

$$\begin{aligned}\Theta_{i,j} &= \frac{i\pi}{H} + \frac{\pi}{2H}, \quad i = \{0,\dots,H-1\},\ j = \{0,\dots,W-1\},\\ \Phi_{i,j} &= -\frac{2\pi j}{W} + \pi - \frac{\pi}{W}, \quad i = \{0,\dots,H-1\},\ j = \{0,\dots,W-1\}.\end{aligned} \tag{35}$$

Here, Θ and Φ are angle matrices of panoramic images with size $H \times W$, which consist of two dimensional Euler angular equivariant series. Given the representation in spherical coordinate systems, each 3D point $(X_{i,j}, Y_{i,j}, Z_{i,j})$ in the camera coordinate system will be obtained through the calculation in Eq. (36).

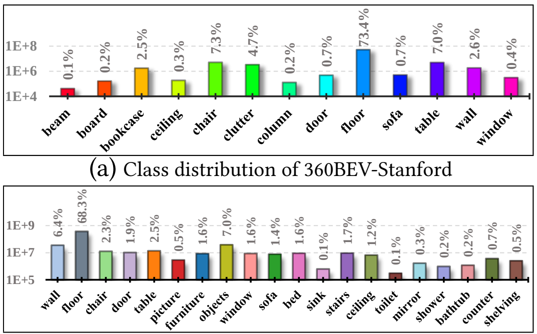
$$\begin{aligned}X_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \sin(\Phi_{i,j}),\\ Y_{i,j} &= D_{i,j} \cdot \cos(\Theta_{i,j}),\\ Z_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \cos(\Phi_{i,j}),\end{aligned} \tag{36}$$

where D is the depth information. After obtaining 3D points, the orthographic projection matrix P_v is applied to transform 3D coordinates to 2D panoramic BEV indices (u, v), as in Eq. (37), where $[\mathbf{R}|\mathbf{t}]$ is the transformation matrix.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{R}^{-1} \begin{bmatrix} X_{i,j} \\ Y_{i,j} \\ Z_{i,j} \end{bmatrix} - \mathbf{t}, \quad \underbrace{\begin{bmatrix} u \\ v \\ 0 \\ 1 \end{bmatrix} = P_v \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}}_{\text{Orthographic projection}}. \tag{37}$$

Dataset	#Scene	#Room	#Frame	#Category
train	5	215	1,040	13
val	1	55	373	13
360BEV-Stanford	6	270	1,413	13
train	61	–	7,829	20
val	7	–	772	20
test	18	–	2,014	20
360BEV-Matterport	86	2,030	10,615	20

Table 22: The statistics of the 360BEV-Matterport and 360BEV-Stanford datasets.



(a) Class distribution of 360BEV-Stanford

(b) Class distribution of 360BEV-Matterport

Figure 47: Per-class pixel number (logarithmic) and frequency (%) distribution.

5.3.2.3 *Dataset Statistics*

As a result, two BEV datasets for panoramic semantic mapping are obtained. The data statistics of 360BEV-Stanford and 360BEV-Matterport datasets are shown in Table 22. While the 360BEV-Stanford dataset has 13 classes and 1,413 images, the 360BEV-Matterport dataset includes 20 classes and 10,615 samples. Besides, we further present the per-class pixel number and per-class frequency in Figure 47. Note that the *floor* class has a much higher frequency on both datasets. This category is important for tasks that rely on complete maps, such as indoor navigation, and is therefore retained.

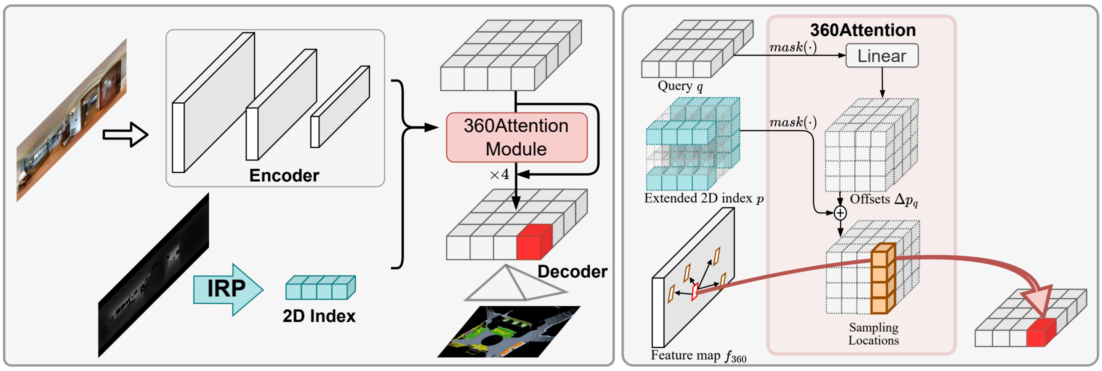
5.3.3 *360Mapper Model*

Overall Architecture. As shown in Figure 48, our end-to-end 360Mapper framework includes four steps: (1) The transformer-based backbone extracts feature from the panoramic image. (2) The *Inverse Radial Projection (IRP)* module obtains a 2D index by projecting reference points generated by depth. (3) The *360Attention* module enhances the front-view feature by 2D index and generates offsets from BEV queries to eliminate the effects of distortion. (4) The lightweight decoder parses the projected feature map and predicts the semantic BEV map.

Inverse Radial Projection. We propose Inverse Radial Projection (IRP), which the input of panoramic depth is included. We can obtain a top-view mask map by using depth information. This mask map is used to generate 3D reference points with the corresponding map height. The 3D reference points are projected onto the sphere to generate 2D reference indexes, as shown in Eq. (38), where ID_h and ID_w represent the index values of 2D reference for the height and width of the feature map, respectively. The 2D indexes are used to locate the corresponding points of front-view feature.

$$\begin{aligned} \Phi &= \tan^{-1}\frac{y}{x}, \quad \Theta = \tan^{-1}\left(\frac{x}{z}\cdot\frac{1}{\cos(\Phi)}\right), \\ \mathrm{ID}_h &= \left\lceil\frac{H\Theta}{\pi}\right\rceil, \quad \mathrm{ID}_w = \left\lceil\left(\frac{\Phi}{\pi}-\frac{1}{W}\right)\cdot\frac{W}{2}\right\rceil. \end{aligned} \tag{38}$$

360Attention. In Figure 48b, the proposed 360Attention generates sampling offsets through the linear layer in an adaptive manner. Given the BEV query $\boldsymbol{q} \in \mathbb{R}^{N\times C_{\mathrm{Emb}}}$ as input, where $N{=}h{\times}w$ is the length of query, a $\mathrm{mask}(\cdot)$ operation is applied on $\boldsymbol{q}$



(a) 360Mapper model

(b) 360Attention module

Figure 48: Architecture of 360Mapper and the 360Attention module. The 360Mapper model includes the encoder for extracting features from the front-view panoramic image, the 360Attention module for feature projection, and the decoder for parsing the projected feature to the BEV semantic map. The offsets are obtained by a linear layer and added with the 2D index that is obtained by Inverse Radial Projection (IRP), yielding the sampling locations for 360BEV feature projection.

and $\boldsymbol{p}$ to mask out irrelevant points and 2D indexes according to the mask map M_{map} from IRP, which is crucial to keep $\boldsymbol{q}$ and $\boldsymbol{p}$ efficient and reducing computation of 360Attention ($\sum M_{map}<N$). The sampling offset $\Delta \boldsymbol{p}_{q,ij}$ and attention weight $\mathcal{A}_{ij} \in [0, 1]$ are predicted through BEV query by linear layers. The adaptive sampling offsets are then added to the extended 2D index $\boldsymbol{p}$ to obtain distortion-aware sampling locations. The 360Attention module can be denoted as:

$$360\mathrm{Attn}(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{f}_{360}) = \sum_{i=1}^{N_{\mathrm{head}}} \mathcal{W}_i \sum_{j=1}^{N_{\mathrm{point}}} \mathcal{A}_{ij} \cdot \boldsymbol{f}_{360} \left(mask\left(\boldsymbol{p}\right) + \Delta \boldsymbol{p}_{q,ij} \right), \tag{39}$$

where $\boldsymbol{q}$, $\boldsymbol{p}$, and $\boldsymbol{f}_{360}$ indicate the query, the extended 2D index, and panoramic feature map, respectively. The linear layer $\mathcal{W}_i \in \mathbb{R}^{C \times (C/N_{head})}$ is specific to each attention head i, where C is the feature dimension and N_{head} is the number of heads. The attention weight $\mathcal{A}_{ij}$ represents the importance of the sampled points j, where $\sum \mathcal{A}_{ij}=1$. The panoramic features $\boldsymbol{f}_{360}$ and the adaptive sampling locations $(mask(\boldsymbol{p})+\Delta \boldsymbol{p}_{q,ij})$ are aggregated using attention weights $\mathcal{A}_{ij}$. Then, the mask map M_{map} is applied to assemble the BEV output as $\boldsymbol{q}' \in \mathbb{R}^{N \times C_{Emb}}$. After being added with a residual term of $\boldsymbol{q}$, the BEV result from $\boldsymbol{q}+\boldsymbol{q}'$ is forwarded to the next 360Attention module.

5.3.4 *Experiments and Analysis*

5.3.4.1 *Panorama Semantic Mapping (360BEV)*

Results on 360BEV-Stanford. In Table 24, SegFormer [239] and SegNeXt [70] are in Early projection mode, which reach unsatisfactory results, which indicate that the pre-projected RGB maintains less rich spatial and visual information of front-view images. Using Late projection, SegFormer with MiT-B2 achieves 18.65% mIoU and surpasses the Early projection, still yielding sub-optimal semantic mapping results. Interestingly, all methods using Intermediate projection obtain >30% mIoU. While using the same MiT-B2 backbone and our 360Mapper achieves 45.78% with +9.70% gains over

Method	Backbone	Acc	mRecall	mPrecision	mIoU
(1) Early projection: Proj.→Enc.→Seg.					
SegFormer [239]	MiT-B2	71.69	20.82	26.34	14.15
SegNeXt [70]	MSCA-B	79.77	34.13	47.39	25.85
(2) Late projection: Enc.→Seg.→Proj.					
HoHoNet [192]	ResNet101	70.01	31.62	30.46	18.49
Trans4PASS [283]	MiT-B2	65.73	31.08	33.15	17.86
Trans4PASS+ [284]	MiT-B2	66.11	38.06	34.14	20.44
SegFormer [239]	MiT-B2	70.50	30.97	30.65	18.65
(3) Intermediate projection: Enc.→Proj.→Seg.					
BEVFormer [111]	MiT-B2	85.50	40.22	51.71	31.69
Trans4Map [25]	MiT-B0	86.41	40.45	57.47	32.26
Trans4Map [25]	MiT-B2	86.53	45.28	62.61	36.08
Trans4Map [25]	MiT-B4	86.99	46.18	58.19	36.69
Ours	MiT-B0	92.07	50.14	65.37	42.42
Ours	MiT-B2	**92.80**	53.56	67.72	45.78
Ours	MSCA-B	92.67	**55.02**	**68.02**	**46.44**

Table 24: Panoramic semantic mapping (360BEV) on the 360BEV-Stanford dataset.

Method	Backbone	Acc	mRecall	mPrecision	mIoU
(1) Early projection: Proj.→Enc.→Seg.					
SegFormer [239]	MiT-B2	68.12	41.33	45.25	29.22
SegNeXt [70]	MSCA-B	68.53	42.13	46.12	30.01
(2) Late projection: Enc.→Seg.→Proj.					
HoHoNet [192]	ResNet101	62.84	38.99	44.22	26.21
Trans4PASS [283]	MiT-B2	55.99	29.59	40.91	20.07
Trans4PASS+ [284]	MiT-B2	57.89	32.75	40.93	21.58
SegFormer [239]	MiT-B2	62.98	41.84	45.30	27.78
(3) Intermediate projection: Enc.→Proj.→Seg.					
BEVFormer [111]	MiT-B2	72.99	43.61	51.70	32.51
Trans4Map [25]	MiT-B0	70.19	44.31	50.39	31.92
Trans4Map [25]	MiT-B2	73.28	51.60	53.02	36.72
Trans4Map [25]	MiT-B4	73.51	50.78	56.67	38.04
Ours	MiT-B0	75.44	48.80	56.01	36.98
Ours	MiT-B2	78.80	59.54	59.97	44.32
Ours	MSCA-B	**78.93**	**60.51**	**62.83**	**46.31**

Table 25: Panoramic semantic mapping (360BEV) on 360BEV-Matterport *val* set.

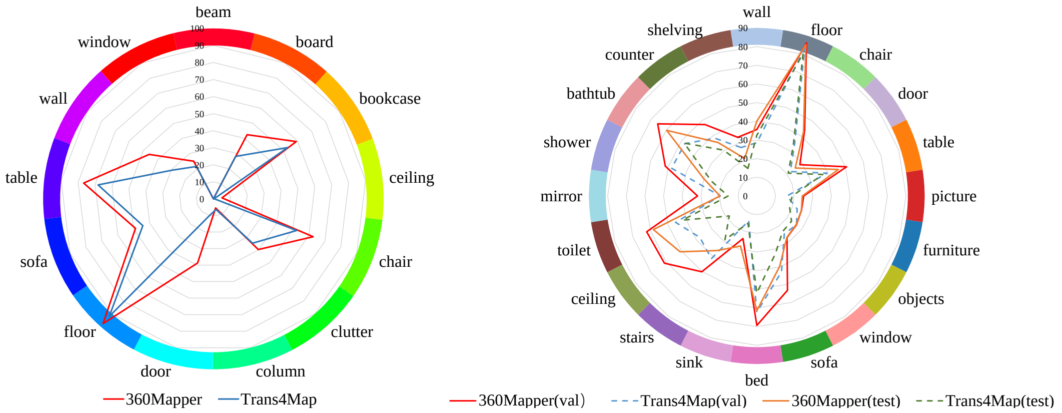
Trans4Map [25]. Further, our model (MiT-B0) outperforms Trans4Map (MiT-B4) with +05.73%. With MSCA-B from SegNeXt [70], our method reaches 46.44% in mIoU, which indicates 360Mapper is flexible to CNN- and Transformer-based backbones.

Results on 360BEV-Matterport. In Table 25, SegFormer [239] and SegNeXt [70] adopt Early projection and show better performance than the Late projection. Using Intermediate projection, our 360Mapper models based on MiT-B0 and MiT-B2, obtain 36.98% and 44.32% in mIoU, respectively. Compared to Trans4Map [25] (MiT-B2), our approach with MiT-B2 has improvements by +5.52% in accuracy, +7.94% in mRecall, +6.95% in mPrecision, and +7.60% in mIoU. Surprisingly, our 360Mapper with MiT-B2 outperforms Trans4Map with MiT-B4 with +6.28% in mIoU. Besides, to compare multi-view methods, we reproduce BEVFormer [111] by using a single panorama instead of six pinhole views. Our 360Mapper has +11.81% gains. Furthermore, we verify 360Mapper by using a CNN-based MSCA-B backbone [70], which obtains 46.31% in mIoU. All results are in line with our observation that Intermediate projection can preserve dense visual cues and long-range information from front-view panoramas.

Per-class Results. We present the comparison of per-class results in Figure 49. Both the baseline Trans4Map and our 360Mapper model are based on the same backbone, *i.e.*, MiT-B2. On the 360BEV-Stanford dataset (Figure 49a), our 360Mapper model has significant gains on most of categories, such as *board* (>14%), *wall* (>16%), *door* (>28%), etc. On the 360BEV-Matterport dataset (Figure 49b), it is readily apparent that our model can better recognize the *chairs* and *tables*, yielding >6% IoU gains compared to Trans4Map [25]. On the test set of the 360BEV-Matterport dataset, our 360Mapper obtains IoU gains with >12% and >15% on the *sink* and *toilet* classes, as compared to Trans4Map. Overall, the consistent improvements on both datasets show the superiority of our 360Mapper on panoramic semantic mapping.

5.3.4.2 *Qualitative Analysis*

To analyze the predicted semantic maps, we visualize the results from the validation set of the 360BEV-Matterport dataset. In Figure 50, from left to right are input images, results of baseline [25], results of our 360Mapper, and ground truth. Thanks to the IRP projection and 360Attention, the segmentation results of 360Mapper are much bet-



(a) 360BEV-Stanford

(b) 360BEV-Matterport

Figure 49: Distribution of per-class semantic mapping results (per-class IoU in %).

ter. In the first scene in Figure 50, 360Mapper is able to successfully classify *chairs*, while the baseline model fails, predicting several *tables* and misclassifying the distant ground as another *table*. In the second scene, the segmentation of the *tables* derived by the baseline is incomplete. Furthermore, in the last zoomed-in scene, 360Mapper provides accurate semantic maps, such as in *counter*, *chair*, and *wall* categories, whereas the baseline Trans4Map [25] misclassifies them as *tables* and *doors*. Based on the qualitative analysis, our 360Mapper can effectively handle object deformations and image distortions, yielding better BEV semantic maps.

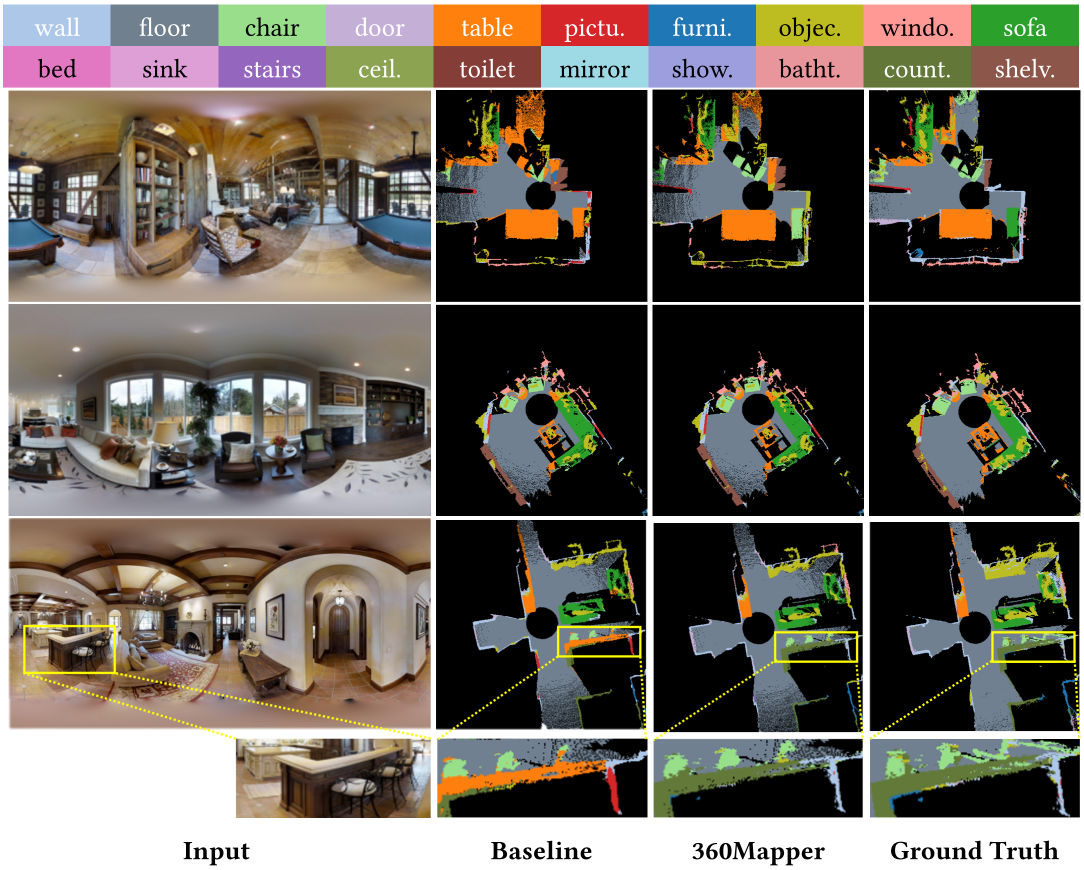


Figure 50: Qualitative analysis on the 360BEV-Matterport dataset. Black regions are void.

5.4 CHAPTER CONCLUSION

Navigational scene understanding is a crucial component of Mobility Assistance Systems (MAS), which can empower the ability of MAS to support independent navigation and provide People with Visual Impairments (PVI) accessible maps for scene exploration prior to visits. As the first research theme in field of MAS, this chapter presents three novel approaches related to navigational scene understanding:

- **Feature matching for visual localization**: Feature matching is a fundamental component of visual localization and navigation systems for PVI. A robust feature matching model is able to stabilize pose estimation and indoor localization performance in low-texture indoor environments.
- **Bird's-eye-view semantic mapping**: Semantic mapping can transform front-view observations into a bird's-eye-view (BEV) semantic map. The pixel-wise semantic map can be used to plan paths in MAS, and provide accessible features for blind people to understand their destination and make plans before visiting.
- **Panoramic semantic mapping**: 360BEV is a novel mapping task, *i.e.*, creating a semantic map of a scene by only using a panoramic image. This task can simplify the mapping process by bypassing the need for complex multi-view perception and narrow-view sequence generation. Besides, it helps to identify objects and landmarks that would be obscured from a single viewpoint.

Here is a more detailed overview of the contributions of each section in this chapter:

Contribution 1: We propose a new *extract-and-match* pipeline that synchronizes feature extraction and feature matching by interleaving self- and cross-attention modules. It is implemented in a novel vision transformer model, *MatchFormer*, which has a robust hierarchical transformer encoder and a lightweight decoder. MatchFormer achieves state-of-the-art results on matching low-texture indoor images, and outperforms previous detector-based and extract-to-match methods.

Contribution 2: We propose an end-to-end Transformer for Mapping (*Trans4Map*), to perform egocentric-to-allocentric semantic mapping. Trans4Map achieves a holistic dense understanding for indoor exploration by capturing long-range contextual dependencies. The novel *Bidirectional Allocentric Memory (BAM)* is also more efficient than previous methods, improving the mapping performance of the model with fewer parameters.

Contribution 3: A new task, *360BEV*, is established for the first time to address indoor semantic mapping using a single-frame panoramic image. It reduces the complex processing of multi-view or sequence inputs. Besides, two indoor BEV datasets, *i.e.*, 360BEV-Matterport and 360BEV-Stanford, are extended with front-view panoramic images and BEV semantic labels, providing a thorough benchmark for panoramic semantic mapping. The 360Mapper model is proposed as a dedicated solution for interior panoramic semantic mapping.

6 TOWARDS GENERALIZABLE SCENE UNDERSTANDING

> In this chapter, we present the second research theme in MAS, *i.e.*, *generalizable scene understanding*. In real-world scenarios, corner cases and adverse situations are challenging for MAS. For example, transparent objects often present architectural barriers which hinder the mobility of People with Visual Impairments (PVI), while adverse driving conditions are difficult for autonomous vehicles. Therefore, scene understanding models should be adaptable to different corner cases. To achieve this, our contributions are two-fold: (1) We propose an efficient Transformer for Transparency (*Trans4Trans*) to recognize glass-like objects, which is presented in Section 6.1, based on our work pulished in *ICCV ACVR workshop 2021* [281]. (2) Considering the synergy between walking and driving scene understanding, Trans4Trans is further verified on driving scenes. Besides, accidental cases are explored through unsupervised domain adaptation. These two methods are presented in Section 6.2, based on our works published in *Transactions on ITS 2022* [282] and *CVPR WAD workshop 2022* [133].

6.1 TRANSPARENT OBJECT SEGMENTATION

This section is based on our work published in *ICCV Workshop on Assistive Computer Vision and Robotics (ACVR) 2021* [281].

6.1.1 *Scene Understanding in the Wild*

Knowledge of glass architecture [16] and glass doors [138, 141] are particular important for People with Visual Impairments (PVI), because transparent objects often present architectural barriers which hinder their mobility. For example, a path behind a glass door is not a free way to navigate (Figure 51) unless it is correctly recognized and reacted. However, most common vision-based navigation assistance systems [2, 216, 251] cannot handle transparent obstacles well, as 3D vision-based methods hardly recover the depth information of texture-less transparent surfaces [2, 251], whereas conventional image segmentation-based methods do not cover the categories of challenging transparent objects [120, 246]. In addition, guide dogs often get confused leading people with blindness to full-pane windows, and differentiation between doors, and large glass windows is difficult for people with residual sight [170]. A system that supports the recognition of landmarks such as doors is particularly appreciated by people with

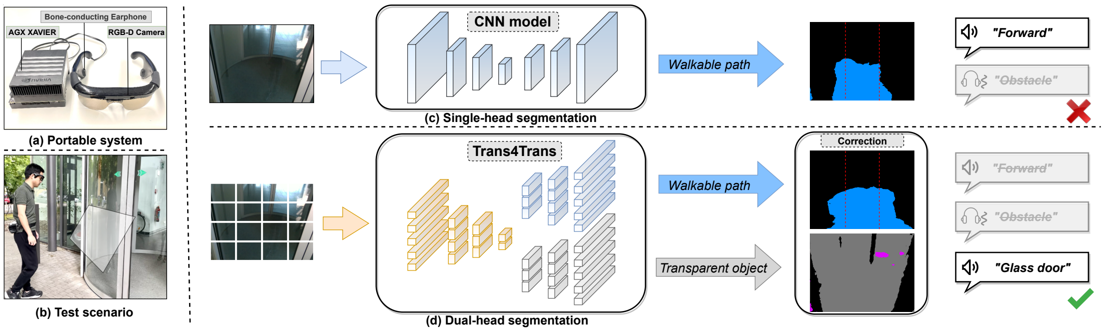


Figure 51: The overview of *Vision4Blind* assistance system. (a) The system equipped with smart vision glasses and a portable GPU is tested (b) in front of a glass door. The input image is segmented as ***Walkable Path*** by (c) a single head CNN model, and is corrected as ***Glass Door*** by (d) our Transformer for Transparency (Trans4Trans) model. The user interface consists of *vibration* and *speech* feedback, such as "*glass door*" in this case.

visual impairments, as finding a door before entering a building is difficult due to the inaccuracy of GPS [10, 170].

To address that, we build a wearable system, *i.e.*, *Vision4Blind*, which can perform real-time wayfinding and object segmentation to assist PVI travel safely. Specifically, we present *Transformer for Transparency (Trans4Trans)*, an efficient semantic segmentation architecture with dual heads, as shown in Figure 51d. As transparent objects are often texture-less or share similar content as the surroundings, it is essential to associate long-range visual concepts to robustly infer transparent regions. For this reason, Trans4Trans is established with both transformer-based encoder and decoder to fully exploit the long-range context modeling capacity of self-attention layers in transformers [212]. In particular, Trans4Trans features a novel *Transformer Paring Module (TPM)* to fuse multi-scale feature maps generated from embeddings of dense partitions, and the symmetric transformer-based decoder can consistently parse the feature maps from transformer-based encoder. Together with predicting general things and stuff classes like walkable areas, the dual-head design can segment transparent objects accurately, which are safety-critical for navigation.

Trans4Trans is integrated in our *Vision4Blind* wearable system which comprises a pair of smart vision glasses and a mobile GPU processor, which delivers a generalizable scene understanding swiftly and accurately thanks to the high efficiency of our model. With the complete semantic information, the user interface consists of vibration and acoustic feedback of detected objects, walkable directions and warnings of the obstacles, which yields intuitive suggestions and no prior knowledge is needed. The system will be described in Section 7.1. Besides, a comprehensive set of experiments has been conducted on multiple semantic segmentation datasets [5, 237]. In particular, the proposed model outperforms state-of-the-art methods on the test sets of Stanford2D3D [5] and Trans10K-v2 [237] datasets. Finally, a user study with visually impaired people and a variety of field tests demonstrate the usability and reliability of our assistive system for navigational perception in the wild. To the best of our knowledge, we are the first to use vision transformers for assisting people with visual impairment.

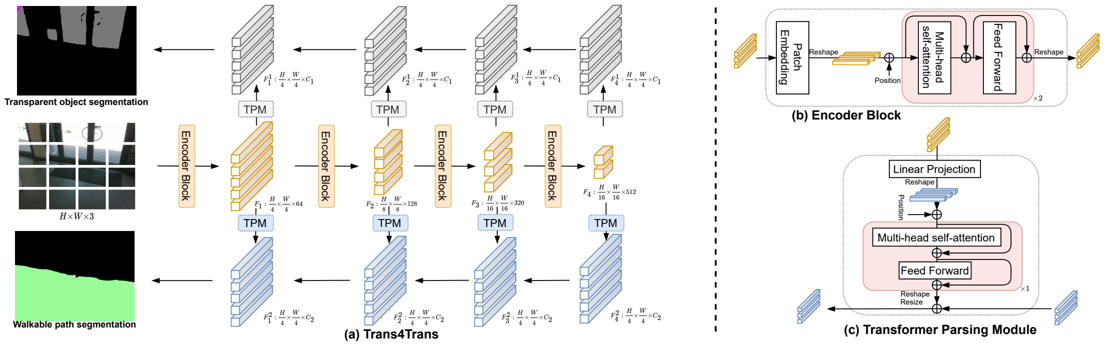

Transparent object segmentation
$H \times W \times 3$
Walkable path segmentation
Encoder Block
Encoder Block
Encoder Block
Encoder Block
TPM
$F_1^1 : \frac{H}{4} \times \frac{W}{4} \times C_1$
$F_2^1 : \frac{H}{4} \times \frac{W}{4} \times C_1$
$F_3^1 : \frac{H}{4} \times \frac{W}{4} \times C_1$
$F_4^1 : \frac{H}{4} \times \frac{W}{4} \times C_1$
$F_1 : \frac{H}{4} \times \frac{W}{4} \times 64$
$F_2 : \frac{H}{8} \times \frac{W}{8} \times 128$
$F_3 : \frac{H}{16} \times \frac{W}{16} \times 320$
$F_4 : \frac{H}{16} \times \frac{W}{16} \times 512$
$F_1^2 : \frac{H}{4} \times \frac{W}{4} \times C_2$
$F_2^2 : \frac{H}{4} \times \frac{W}{4} \times C_2$
$F_3^2 : \frac{H}{4} \times \frac{W}{4} \times C_2$
$F_4^2 : \frac{H}{4} \times \frac{W}{4} \times C_2$
(a) Trans4Trans
Patch Embedding
Reshape
Position
Multi-head self-attention
Feed Forward
Reshape
×2
(b) Encoder Block
Linear Projection
Reshape
Position
Multi-head self-attention
Feed Forward
×1
Reshape
Resize
(c) Transformer Parsing Module

Figure 52: The architecture of (a) Trans4Trans model consists of shared encoder and dual decoders, while (b) and (c) are the general transformer-based encoder block and our proposed Transformer Parsing Module (TPM) for decoder, respectively.

6.1.2 *Trans4Trans Model*

Inspired by the benefit of ViT [52] transformer model in acquiring long-range dependencies, our dual-head Trans4Trans model is entirely composed of transformers, as shown in Figure 52a, while the single-head has only one decoder. The four-stage encoder is borrowed from PVT [222]. Different to PVT-based Trans2Seg [237] adopting CNN-decoder, both encoder and decoder of Trans4Trans are symmetrically constructed by transformers for maintaining consistency in both feature extraction and feature parsing stages. Furthermore, different from CNN-based models [140, 155, 258, 263] learning the inductive bias, the transformer-based decoder is supposed to be more robust to parse unseen data captured in the wild. Yet, training a transformer model requires a large-scale dataset [52]. In order to solve the data-hunger problem and correct the misidentified walkable area through transparent objects segmentation, we designed a double-head model. Through the joint training of multiple datasets, it brings greater data diversity for learning a robust transformer-based model.

To construct a lightweight decoder, we propose a *Transformer Parsing Module (TPM)* as shown in Figure 52c. Each TPM contains one single transformer-based layer, thus it is flexible to be deployed on our portable hardware system. Precisely, each stage has a TPM module and contains similar structure. As shown in Figure 52a, the pyramid features $\{F_1, F_2, F_3, F_4\}$ from encoder are parsed consistently by the specific TPM module. Between two stages, resize and element-wise addition are used for pyramid feature fusion. For balancing capacity and computational demands, the feature resolution of each TPM is set as $\frac{H}{4} \times \frac{W}{4} \times C$, for which the default channel is 64.

Thanks to TPM, the amount of GFLOPs and parameters of this dual-head structure is largely reduced compared to two separate models. Besides, diverse features can be learned from various datasets. Thereby, the dual-head model maintains lightweight and is robust in terms of preventing overfitting when testing in real-world scenarios. The decoder composed of our TPM module can be flexibly applied with various CNN- or transformer-based encoder structures as well. For multi-task learning, mounting decoder heads robustifies the feature learned via the shared encoder, and the entire model will not be computationally overburdened.

Method	GFLOPs↓	ACC↑	mIoU↑	Category IoU ↑											
				Backg.	Shelf	Jar/Tank	Freezer	Window	Door	Eyeglass	Cup	Wall	Bowl	Bottle	Box
FPENet [123]	0.76	70.31	10.14	74.97	0.01	0.00	0.02	2.11	2.83	0.00	16.84	24.81	0.00	0.04	0.00
ESPNetv2 [140]	0.83	73.03	12.27	78.98	0.00	0.00	0.00	0.00	6.17	0.00	30.65	37.03	0.00	0.00	0.00
ContextNet [154]	0.87	86.75	46.69	89.86	23.22	34.88	32.34	44.24	42.25	50.36	65.23	60.00	43.88	53.81	20.17
FastSCNN [155]	1.01	88.05	51.93	90.64	32.76	41.12	47.28	47.47	44.64	48.99	67.88	63.80	55.08	58.86	24.65
DFANet [103]	1.02	85.15	42.54	88.49	26.65	27.84	28.94	46.27	39.47	33.06	58.87	59.45	43.22	44.87	13.37
ENet [151]	2.09	71.67	8.50	79.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.25	0.00	0.00	0.00
HRNet_w18 [217]	4.20	89.58	54.25	92.47	27.66	45.08	40.53	45.66	45.00	68.05	73.24	64.86	52.85	62.52	33.02
HarDNet [24]	4.42	90.19	56.19	92.87	34.62	47.50	42.40	49.78	49.19	62.33	72.93	68.32	58.14	65.33	30.90
DABNet [100]	5.18	77.43	15.27	81.19	0.00	0.09	0.00	4.10	10.49	0.00	36.18	42.83	0.00	8.30	0.00
LEDNet [226]	6.23	86.07	46.40	88.59	28.13	36.72	32.45	43.77	38.55	41.51	64.19	60.05	42.40	53.12	27.29
Trans4Trans-T	10.45	93.23	68.63	94.44	48.39	61.89	61.86	61.14	54.83	73.60	83.03	75.20	74.69	75.26	59.19
ICNet [292]	10.64	78.23	23.39	83.29	2.96	4.91	9.33	19.24	15.35	24.11	44.54	41.49	7.58	27.47	3.80
BiSeNet [258]	19.91	89.13	58.40	90.12	39.54	53.71	50.90	46.95	44.68	64.32	72.86	63.57	61.38	67.88	44.85
Trans4Trans-S	19.92	94.57	74.15	95.60	**57.05**	71.18	**70.21**	63.95	61.25	81.67	87.34	78.52	**77.13**	81.00	64.88
DenseASPP [253]	36.20	90.86	63.01	91.39	42.41	60.93	64.75	48.97	51.40	65.72	75.64	67.93	67.03	70.26	49.64
DeepLabv3+ [29]	37.98	92.75	68.87	93.82	51.29	64.65	65.71	55.26	57.19	77.06	81.89	72.64	70.81	77.44	58.63
FCN [127]	42.23	91.65	62.75	93.62	38.84	56.05	58.76	46.91	50.74	82.56	78.71	68.78	57.87	73.66	46.54
OCNet [263]	43.31	92.03	66.31	93.12	41.47	63.54	60.05	54.10	51.01	79.57	81.95	69.40	68.44	78.41	54.65
RefineNet [116]	44.56	87.99	58.18	90.63	30.62	53.17	55.95	42.72	46.59	70.85	76.01	62.91	57.05	70.34	41.32
Trans2Seg [237]	49.03	94.14	72.15	95.35	53.43	67.82	64.20	59.64	60.56	**88.52**	86.67	75.99	73.98	82.43	57.17
TransLab [240]	61.31	92.67	69.00	93.90	54.36	64.48	65.14	54.58	57.72	79.85	81.61	72.82	69.63	77.50	56.43
DUNet [89]	123.69	90.67	59.01	93.07	34.20	50.95	54.96	43.19	45.05	79.80	76.07	65.29	54.33	68.57	42.64
U-Net [166]	124.55	81.90	29.23	86.34	8.76	15.18	19.02	27.13	24.73	17.26	53.40	47.36	11.97	37.79	1.77
DANet [61]	198.00	92.70	68.81	93.69	47.69	66.05	70.18	53.01	56.15	77.73	82.89	72.24	72.18	77.87	56.06
PSPNet [293]	187.03	92.47	68.23	93.62	50.33	64.24	70.19	51.51	55.27	79.27	81.93	71.95	68.91	77.13	54.43
Trans4Trans-M	34.38	**95.01**	**75.14**	**96.08**	55.81	**71.46**	69.25	**65.16**	**63.96**	83.84	**88.21**	**80.29**	76.33	**83.09**	**68.09**

Table 26: Computation complexity in GFLOPs and category-wise accuracy evaluation and comparison with semantic segmentation methods on the Trans10K-v2 dataset [237].

6.1.3 *Experiments and Analysis*

6.1.3.1 *Comparison to Advanced Models.*

Following [237], we compare accuracy- and efficiency-oriented models as shown in Table 26. Compared with both CNNs and transformer-based methods like Trans2Seg [237], the superiority of Trans4Trans is further confirmed. Our Trans4Trans-M model outperforms the advanced method Trans2Seg by 2.99% in mIoU and 0.87% in ACC, while requiring much less GFLOPs. For category-wise accuracy, our Trans4Trans model achieves the advanced IoU on the classes *background*, *jar or tank*, *window*, *door*, *cup*, *wall*, *bottle* and *box*. These experimental results show the efficacy of transparent object segmentation of the proposed Trans4Trans architecture.

6.1.3.2 *Real-time Performance*

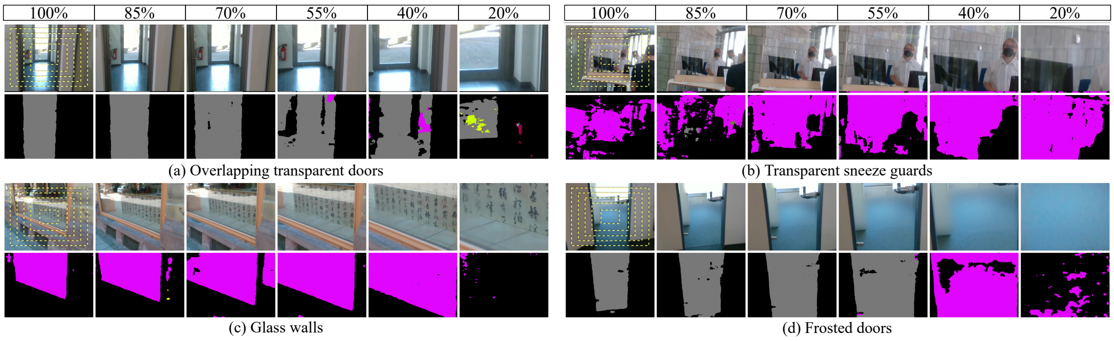
To calculate the inference speed of our different versions of dual-head Trans4Trans model, 300 samples from the Trans10K-v2 test set with a batch size of 1 and a resolution of 512×512 are tested on three different GPUs, *i.e.*, a mobile NVIDIA AGX Xavier in the MAXN mode, an NVIDIA GeForce MX350 from a lightweight laptop and an RTX 2070 from a workstation. As shown in Table 27, the computation costs of our tiny Trans4Trans model on three GPUs are considerably lower than the other two, meanwhile the performances of the three models on both datasets are suitable for our system. In real applications, the more timely response of the navigation system is beneficial for assisting users with a similar prediction accuracy on each frame. Hence, the tiny version is selected in the user study.

Network	NVIDIA Xavier (ms) ↓	MX350 (ms) ↓	RTX 2070 (ms) ↓
Trans4Trans-M	115.9 (±1.1) / 202.8 (±1.1)	186.1 (±0.3) / 243.2 (±0.3)	22.9 (±0.3) / 36.6 (±0.8)
Trans4Trans-S	95.3 (±0.6) / 158.6 (±1.8)	140.6 (±0.3) / 188.4 (±0.4)	17.1 (±0.3) / 27.7 (±0.5)
Trans4Trans-T	75.8 (±0.7) / 122.7 (±0.7)	101.5 (±0.3) / 141.7 (±1.6)	12.8 (±0.5) / 20.3 (±0.5)

Table 27: Inference time (ms/frame) of dual-head Trans4Trans is tested in half-/single-precision on various GPUs at 512×512.

6.1.3.3 *Transparent Feature*

To investigate the impact of context information (such as door frames and walls) or reflection features on transparency perception, we visualize the segmentation results from six scales (from 100% to 20%) of the original images in Figure 53. In Figure 53a, among all six scales, even in the 20% case with less context, two overlapping doors are accurately segmented. In Figure 53b, the sneeze guard is recognized as the glass wall, since it has the glass-like reflection. Another reason is the sneeze guard has no frame and overlaps with the background wall. Therefore, the background of transparent objects affects the object classification if they lack contextual information. In the 40% ratio in Figure 53c, with only a one-side outer frame and part of the reflection, it can still segment the area of the glass wall. However, in the 20% ratio, it is confused due to the tiny frame and absence of any reflections. The errors in the latter two ratios of Figure 53d are caused by the lack of texture information and reflections. Based on the analysis of visualizations, three insights are provided: (1) The contextual information, *e.g.*, the outer frame, is a vital factor for the transparency segmentation; (2) The reflection characteristic of glass or transparent objects is crucial; (3) The background texture of transparent objects also interferes with the segmentation results when they lack contextual information. Thanks to the symmetrical encoder-decoder structure, Trans4Trans can robustly segment transparent objects even with diminishing context cues in most of the complex real-life scenes.



(a) Overlapping transparent doors

(b) Transparent sneeze guards

(c) Glass walls

(d) Frosted doors

Figure 53: Visualization of segmentation results from different cropped regions based on image center. Images of six scales from 100% to 20% are cropped from its original images and are separately segmented, in order to ablate the effect of image context. All results are generated by the tiny version of both networks.

6.2 ADVERSE SCENE SEGMENTATION

This section is based on our work published in *Transactions on ITS 2022* [282], and partially from a collaborative work of a master's thesis project published in *CVPR Workshop on Autonomous Driving (WAD) 2022* [133].

6.2.1 *Synergy of Walking and Driving Scene Understanding*

Assisted navigation of pedestrians and automated driving of intelligent vehicles are inextricably intertwined in the Intelligent Transportation Systems (ITS) field [21, 136, 189, 246], both with the aim to improve traffic flow towards the utopia of all road participants. In addition to vehicles from the driving perspective, humans and their mobilities from the walking perspective are involved. However, people with disabilities may have difficulties in using transportation infrastructures, and the bottleneck of inclusiveness should be broken in ITS. To this end, it is necessary to expand the coverage of assistance systems from drivers to pedestrians, especially those with visual impairments, who are one of the most vulnerable road users [137].

To assist People with Visual Impairments (PVI) to navigate, it is essential to attain *adaptable* scene understanding in the *walking* perspective which is verified via the *Vision4Blind* system [281]. However, it shares similar challenges with the ITS research line on *driving* surrounding segmentation [164, 249], when considering the synergy towards traffic safety and the shared challenges between walking and driving scene understanding (Figure 54). Apart from walking scenes, Trans4Trans is further adapted to and verified on driving benchmarks including Cityscapes [44], ACDC [171], and DADA-seg [285]. In the following, we mainly advocate addressing semantic segmentation of driving scenes from an adaptable perspective that jointly considers normal, adverse, and accidental scenarios.

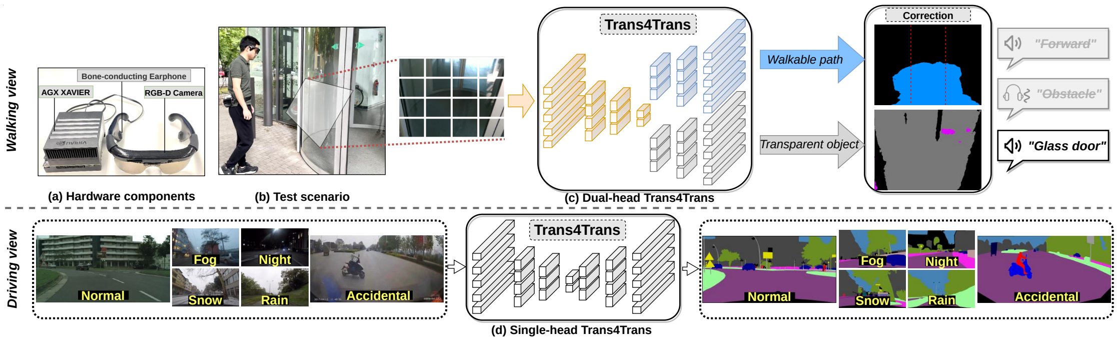


Figure 54: The overview of *Vision4Blind* assistance system in both walking and driving perspectives. (a) The system equipped with the smart vision glasses and a portable processor is tested (b) in front of a glass door. The input image is segmented as ***walkable path*** and ***glass door*** by (c) our Trans4Trans model. The user interface has vibration and voice feedback. After training on normal and adverse data from street scenes, (d) Trans4Trans reaches high robustness in various real-world driving cases, *e.g.*, normal, adverse, and accidental scenes.

Network	Encoder	Decoder	GFLOPs	#P(M)	Cityscapes	ACDC
PVT-T	PVT-T	MiT [237]	10.30	13.11	58.09	53.65
PVT-S	PVT-S		19.77	24.35	59.68	57.13
PVT-M	PVT-M		36.87	51.83	60.38	58.60
Ours-T	PVT-T	TPM / Single-head	10.45	12.71	60.41(+2.32)	54.37(+0.72)
Ours-S	PVT-S		21.98	25.00	63.08(+3.40)	60.70(+3,57)
Ours-M	PVT-M		44.38	48.77	65.63(+5,25)	61.91(+3,31)
Ours-T	PVT-T	TPM / Dual-head	11.23	13.10	57.42(-0.67)	56.36(+2.71)
Ours-S	PVT-S		24.82	26.45	62.39(+2.71)	62.14(+5.01)
Ours-M	PVT-M		55.16	54.28	63.00(+2.62)	63.88(+5.28)
Ours-T	PVTv2-B1	TPM / Single-head	9.18	13.53	63.25(+5.16)	59.25(+5.60)
Ours-S	PVTv2-B2		19.27	25.62	67.28(+7.60)	64.61(+7.48)
Ours-M	PVTv2-B3		41.89	49.55	69.34(+8.96)	65.92(+7.32)
Ours-T	PVTv2-B1	TPM / Dual-head	10.00	13.93	62.31(+4.22)	61.86(+8.21)
Ours-S	PVTv2-B2		22.17	27.08	65.98(+6.30)	64.83(+7.70)
Ours-M	PVTv2-B3		52.77	55.09	69.05(+8.67)	66.65(+8.05)

Table 28: Ablation on Cityscapes and ACDC. GFLOPs at 512×512. The dimension of *model*-T/-S/-M decoder is {64, 128, 256}.

Methods	Encoder	GFLOPs	#P(M)	mIoU
FastSCNN [155]	Fast-SCNN	2.07	1.46	72.65
CGNet [230]	CGNet-M3N21	7.72	0.50	64.80
Trans4Trans-T	PVTv2-B1 [221]	20.66	13.53	78.23
HRNet [217]	HRNetV2p-W18s	21.70	3.94	77.48
SegFormer-B1 [238]	MiT-B1	29.85	13.66	78.43
ERFNet [164]	ERFNet	30.22	2.07	72.10
Trans4Trans-S	PVTv2-B2 [221]	43.37	25.62	80.02
PSPNet [293]	MobileNetV2	119.09	13.72	70.20
PSPNet [293]	ResNet-18	119.27	12.77	76.90
SegFormer-B2 [238]	MiT-B2	127.86	27.33	80.46
SegFormer-B3 [238]	MiT-B3	160.78	47.18	81.50
DeepLabv3+ [29]	MobileNetv2	169.53	18.70	75.20
HRNet [217]	HRNetV2p-W48	210.57	65.86	80.72
PSPNet [293]	ResNet-50	401.51	48.98	79.96
PSPNet [255]	ResNet-101	573.48	67.95	80.04
SETR-Naive [296]	ViT-L [52]	698.52	306.58	77.90
SETR-MLA [296]	ViT-L [52]	712.76	310.81	77.24
SETR-PUP [296]	ViT-L [52]	818.26	319.11	79.34
Trans4Trans-M	PVTv2-B3 [221]	94.25	49.55	**81.54**

Table 29: Comparison on Cityscapes with multi-scale testing. GFLOPs at 768×768.

6.2.2 *Evaluation on Normal and Abnormal Driving Scenes*

6.2.2.1 *Ablation of Trans4Trans*

We first conduct ablation study of Trans4Trans on driving scene datasets. Five groups of results are shown in Table 28. Our TPM-based Trans4Trans illustrates better performance compared to PVT on both driving scene datasets. On Cityscapes, Trans4Trans-M leveraging PVT encoder outperforms PVT-M by 5.25% and Trans4Trans-M leveraging PVTv2 as the encoder surpasses by 8.96%. On ACDC, our Trans4Trans-M with PVT outperforms PVT-M by 5.28% and the one with PVTv2-M exceeds by 8.05% while utilizing TPM/Dual-head in the decoder architecture. Since ACDC has adverse conditions, these results evidence that TPM/Dual-head has the better robustness under environment changes in driving scene segmentation, as it incorporates more generalized knowledge learned from diverse images in both datasets.

6.2.2.2 *Segmentation in Normal Conditions*

In Table 29, results of Trans4Trans trained with the input size of 768×768 are compared with more than 15 state-of-the-art methods[1]. All Trans4Trans are constructed with the best single-head Trans4Trans. Our Trans4Trans-M approach with PVTv2-B3 as encoder achieves the best performance with an mIoU of 81.54% on Cityscapes, whose images are collected under normal weather and favorable illumination conditions. Compared with the advanced methods such as SETR [296] and PSPNet [293], our Trans4Trans approach shows smaller GFLOPs (94.25) and less parameters (49.55M), which are relevant for fast inference in automated vehicles. Trans4Trans-T and -S models with lighter encoder architectures also show high scores of 78.23% and 80.02% in

[1] For a fair comparison, model weights are obtained by the same framework MMSegmentation: https://github.com/open-mmlab/mmsegmentation.

Method	Trained on	GFLOPs ↓	Fog	Night	Rain	Snow	All-ACDC	All-DADA
DeepLabv3+ [29]	CS	178.1	45.7	25.0	50.0	42.0	41.6	10.4
HRNet [217]	CS	210.5	38.4	20.6	44.8	35.1	35.3	15.5
Trans4Trans-M	CS	41.8	**74.1**	**31.1**	**63.4**	**57.9**	**55.7**	**27.7**
DeepLabv3+ [29]	ACDC	178.1	69.1	60.9	74.1	69.6	70.5	26.8
HRNet [217]	ACDC	210.5	74.7	**65.3**	**77.7**	76.3	75.0	27.5
Trans4Trans-M	ACDC	41.8	**79.8**	55.3	77.4	**78.6**	**75.2**	**32.4**
Trans4Trans-M	ACDC+CS	41.8	**81.4**	56.0	77.0	**78.8**	**76.3**	**39.2**

Table 30: Comparison on adverse (Fog, Night, Rain, Snow, and All-ACDC [171]) and accidental (All-DADA [285]) conditions. CS: Cityscapes [44]. GFLOPs are calculated at 768×768.

mIoU. The lightest Trans4Trans outperforms FastSCNN [155] and CGNet [230] by large margins, and it achieves a similar score as SegFormer [238] while being more efficient.

6.2.2.3 *Segmentation in Adverse Conditions*

In Table 30, we adapt and test Trans4Trans-M on both ACDC [171] and DADA-seg [285] datasets, which have adverse- and accidental scenes, respectively. The results of Trans4Trans are obtained via MMSegmentation with a resolution of 768×768. In the first group, Trans4Trans-M obtains 55.7% and 27.7% in mIoU when compared with HRNet [217] and DeepLabV3+ [29]. Trans4Trans outperforms them in all four adverse conditions and accidental scenes, which demonstrates its high adaptation capacity to unseen domains. This is because with both transformer-based encoder and decoder, Trans4Trans can associate long-range visual concepts for robustly inferring semantics, despite local texture- and illumination changes in different scenarios like night-time and accident scenes. Thanks to our efficient backbone, Trans4Trans surpasses HRNet by >20% and >12% on All-ACDC and All-DADA with only its 20% GFLOPs. In the second group of Table 30, Trans4Trans again indicates better overall performances on two datasets. Finally, the model trained on ACDC and Cityscapes shows the best overall scores on All-ACDC and All-DADA with 76.3% and 39.2% in mIoU, illustrating that co-training on normal and adverse data can improve the performance of the model under both adverse and extreme accident conditions.

6.2.2.4 *Visualization of Driving Scene*

In Figure 55, we visualize the predictions of Trans4Trans* trained on Cityscapes and ACDC, in comparison to DeepLabv3+ [29], HRNet [217], and our Trans4Trans models only trained on ACDC. DeepLabv3+ and HRNet produce noisy results in complex conditions, like the *cars* in shadow (the first row). In adverse weather and week illumination conditions, previous methods yield less precise and even fragmented semantics, like the *trucks* in foggy and rainy scenes (the second and fourth rows) and the *sidewalks* in night and snowy scenes (the third and fifth rows). In accident scenes, which are safety-critical for automated vehicles, existing models cannot generate reliable predictions to be propagated to upper-level applications, as the close *pedestrian* is even completely recognized as *road*. In contrast, Trans4Trans, which learns to gather long-range dependency from the very first layers, delivers more robust segmentation in various scenes, as it is less affected by local texture and illumination changes. Trans4Trans

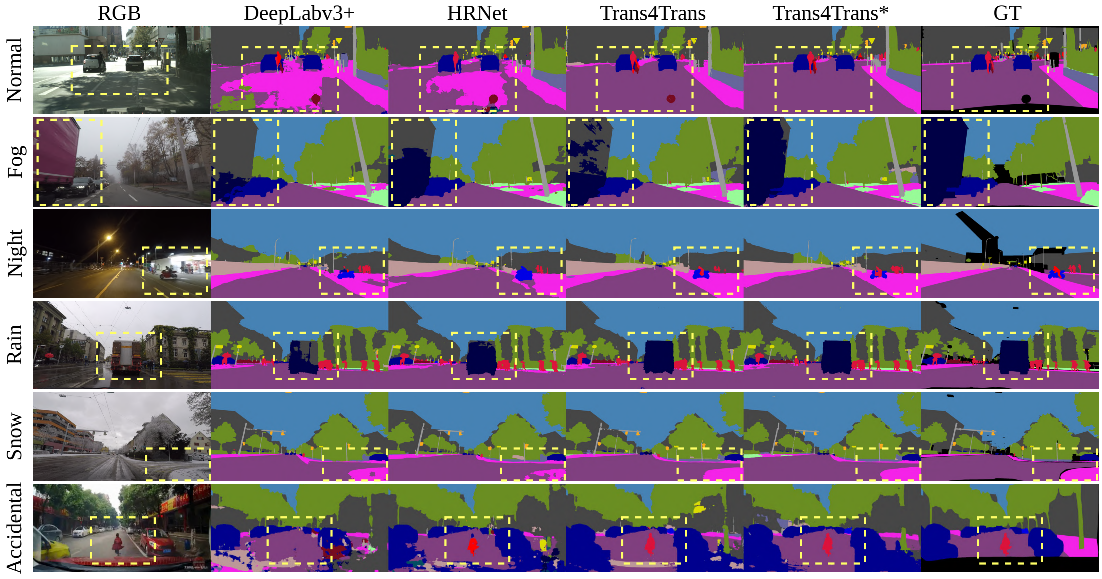


Figure 55: Qualitative analysis on Cityscapes [44] (*Normal*), ACDC [171] (*Fog, Night, Rain*, and *Snow*), and DADA-seg [285] (*Accidental*). The Trans4Trans* is trained on ACDC+Cityscapes, whereas other models are trained on ACDC dataset.

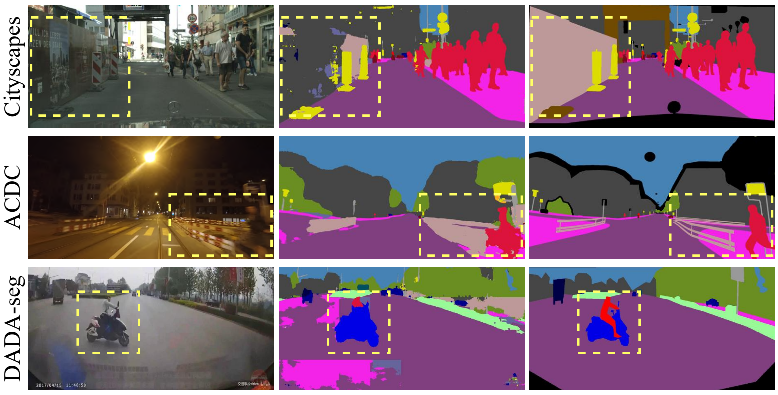


Figure 56: Failure analysis of driving scene segmentation. From left to right are RGB images, segmentation results, and the ground truth.

trained on both adverse and normal datasets further improves the performance, resulting sharp and fine-grained semantic segmentation.

Apart from comparing positive predictions, Figure 56 shows some erroneous semantic segmentation results of the Trans4Trans* model on all three driving datasets. In the first row, the model accurately segments the *pedestrian* on the right side, but the segmentation of the *fence* is less complete because the *fence* is thin and has a similar color to the background. In the second row, the model struggles with the extreme motion blur caused by the dynamic driving of the ego car under very weak illuminations. In the third row, the model fails to segment the abnormal behavior of the *motorcyclist* before the accident. These bad driving cases are very common in real-world autonomous driving, but it is still difficult to identify and deal with them very accurately. One potential solution is to fuse complementary information from different modalities, such as depth, thermal, and event-based sensors [278, 279, 285].

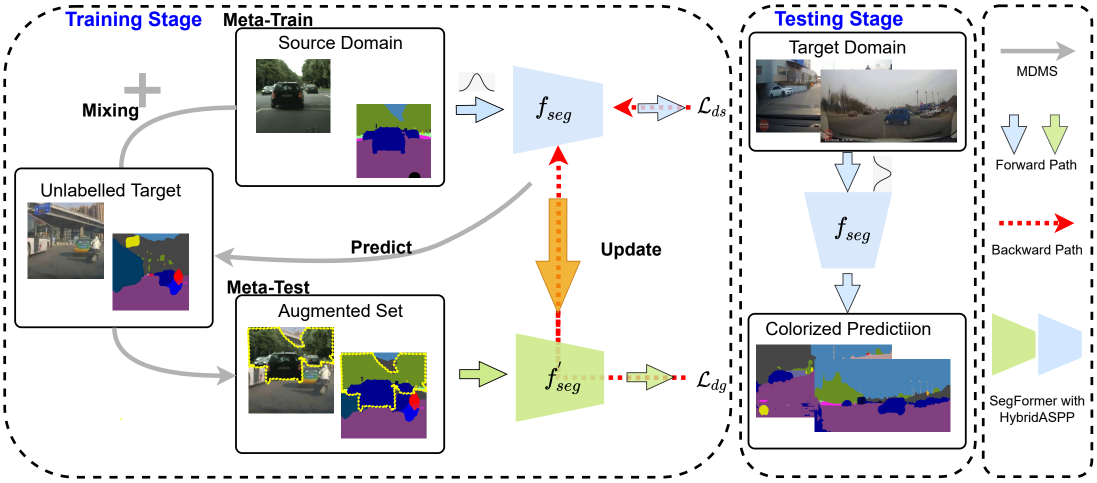


Figure 57: Overview of the *Multi-source Meta-learning UDA (MMUDA)* framework. It includes *Multi-domain Mixed Sampling (MDMS)* and meta-learning with segmentation transformers. Given multiple source (normal) domains, the model fine-tuned by meta-training and meta-testing across various source domains, can generalize well in the target (abnormal) domain.

6.2.3 *Multi-source Meta-learning Adaptation*

To achieve generalizable scene understanding on adverse scenarios, we delve deeper into corner cases within traffic scenes. Our focus lies in the segmentation of *accidental* scenes, as unexpected objects or traffic scenarios constitute a common cause of dangerous situations. A significant portion of real-life accidents features unusual scenes, such as those with object deformations, overturns, and unexpected traffic behaviors.

To enhance the robustness of model in accident scenarios, we introduce a novel approach, *i.e.*, *Multi-source Meta-learning UDA (MMUDA)* framework (Figure 57). To effectively learn from the entirety of the unlabelled target domain dataset, we propose a *Multi-Domain Mixed Sampling (MDMS)* strategy. A source domain is a set of image and label pairs $\{(X_S^i, Y_S^i)\}^{N_S} \in \mathcal{D}_S$, where $X_S^i \in \mathbb{R}^{H \times W \times 3}$ is the image, $Y_S^i \in \mathbb{R}^{H \times W \times C}$ is the C-class label, and N_S is the number of samples in the source domain $\mathcal{D}_S$. From the target domain $\mathcal{D}_T$ with a number of $N_T = N_L + N_U$ samples, the N_U unlabelled image and pseudo label pairs $\{(X_T^i, \hat{Y}_T^i)\}^{N_U} \in \mathcal{D}_T$ are selected for the mixing approach, where $\hat{Y}_T$ is generated by the segmentation transformer f_{seg} in Figure 57. The labelled N_L images in the target domain are only used in the testing stage. In the augmented set $\mathcal{D}_M$ with the same N_U samples, an augmented image X_M is generated by mixing a source image X_S and a target image X_T, and the pseudo label $\hat{Y}_M$ by combining the corresponding ground-truth label Y_S and the pseudo label $\hat{Y}_T$. However, in our case, there are K source domains, thus the augmented set is created as $\{(X_{M_k}^i, \hat{Y}_{M_k}^i)\}^{N_U} \in \mathcal{D}_M$. The cross-entropy is the loss function $\mathcal{L}$ for our task. First, the domain-specific loss $\mathcal{L}_{ds}$ is computed from the meta-training data though f_{seg}. The gradient $\nabla \mathcal{L}_{ds}$ is used to update a new network f'_{seg}, *i.e.*, the green block in Figure 57, which shares with the blue one. To perform generalizable scene understanding in the unseen target domain, the adaptation loss $\mathcal{L}_{da}$ is calculated from f'_{seg} with the updated parameters using the meta-test data. Finally, we employ the total loss $\mathcal{L}_{total} = \mathcal{L}_{da} + \alpha \mathcal{L}_{ds}$ to update the original f_{seg}, optimizing to both source and target domains.

	Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Source-only	MobileNetV2 [172]	16.05	31.87	8.50	26.55	3.60	5.38	13.96	19.51	10.87	44.99	11.09	67.05	8.11	5.23	28.58	11.77	2.17	-	1.90	3.86
	PSPNet [293]	17.07	31.62	11.42	32.48	4.16	8.52	12.38	17.93	13.39	50.82	13.85	67.19	9.86	3.13	31.54	6.97	3.15	-	2.97	2.89
	ResNet50 [74]	18.96	34.19	8.24	31.05	4.56	7.39	19.04	27.05	15.35	33.30	12.40	61.52	10.04	3.95	42.59	14.15	27.02	-	3.72	4.72
	SemFPN [97]	19.59	37.90	10.12	23.80	3.74	9.64	22.06	28.64	15.55	40.95	12.13	51.93	9.24	5.93	52.08	13.89	26.54	-	3.66	4.36
	DNLNet [256]	19.72	41.68	13.26	30.45	6.17	11.04	21.91	28.03	17.99	40.05	14.13	56.06	10.75	5.41	34.78	8.01	28.01	-	3.55	3.39
	ResNeSt [275]	19.99	39.63	11.38	33.68	2.81	9.73	22.76	27.35	18.09	45.24	14.22	71.23	13.34	5.03	36.45	6.91	13.08	-	3.94	4.87
	DANet [61]	22.24	46.49	10.17	42.20	3.81	10.65	13.46	18.69	22.59	55.76	22.22	83.84	6.68	11.75	39.59	7.96	12.64	-	7.98	6.12
	ResNet101 [74]	23.60	57.96	11.16	39.94	6.43	9.46	23.67	27.37	17.32	45.65	16.47	69.21	13.19	4.51	47.29	13.75	30.44	-	6.64	8.01
	OCRNet [262]	24.85	42.13	11.54	34.49	6.63	12.70	22.76	29.03	22.28	42.41	15.15	85.43	14.31	6.65	53.94	20.65	34.86	-	9.30	7.87
	FastSCNN [155]	26.32	69.91	16.30	52.53	6.09	9.63	19.98	19.30	22.58	57.04	22.95	90.81	11.19	13.95	46.16	22.65	9.74	-	4.49	4.75
Cross-source	CLAN [130]	28.76	79.80	18.61	51.56	8.32	13.60	15.51	17.15	21.51	63.20	21.99	80.53	8.37	6.32	63.47	33.43	33.12	-	3.69	6.21
	BDL [109]	29.66	81.44	19.18	57.18	8.61	16.26	14.65	8.78	16.77	66.60	26.83	85.87	10.51	7.16	65.45	35.18	34.78	-	2.71	5.57
	ISSAFE [285]	29.97	80.23	19.51	52.02	6.43	14.68	16.19	17.03	19.50	65.39	21.69	79.84	9.95	8.82	65.60	39.51	39.73	-	6.09	7.03
	EDCNet [286]	32.04	73.03	19.47	57.31	11.60	14.30	20.70	12.27	27.22	70.54	18.98	88.64	10.69	8.70	68.14	49.80	50.86	-	9.02	4.12
	Trans4Trans-M [281]	39.20	71.10	15.57	70.39	10.34	16.53	31.63	**37.16**	**37.38**	71.88	19.61	93.04	21.27	14.97	64.04	53.76	81.53	-	24.63	10.07
	Our Baseline	40.73	84.93	23.66	68.34	16.27	20.58	25.96	31.25	28.20	71.89	22.39	93.16	17.92	26.84	73.89	55.09	69.26	-	34.77	9.49
	+Meta	45.03	86.20	25.44	70.63	14.21	19.75	26.56	28.01	29.23	74.45	25.29	93.18	20.40	31.53	75.02	64.73	76.84	-	38.05	10.95
	+Meta+MDMS	46.11	87.10	27.71	71.11	**22.94**	**20.64**	**32.25**	29.49	34.34	75.48	24.02	92.18	20.65	**33.33**	74.64	63.35	71.14	-	**39.08**	**12.04**
	Our MMUDA	**46.97**	**87.51**	**27.97**	**74.76**	16.16	21.93	29.94	29.43	31.62	**75.67**	**26.69**	**93.57**	**24.40**	29.57	**77.35**	**68.24**	**84.02**	-	36.96	10.44

Table 31: Comparison of state-of-art methods on DADA-seg dataset. The source-only models are trained on the Cityscapes dataset, while the other models are domain-transferred using a single source [109, 130], multiple sources [281, 286] or a different modality [285].

6.2.3.1 *Results of Accident Scenes Segmentation*

Table 31 presents a comparison of mIoU and per-class IoU scores achieved on the DADA-seg dataset [285]. Notably, models trained solely on the Cityscapes dataset experience significant performance degradation and exhibit relatively low accuracy when applied to abnormal accident scenes. For instance, the source-only ResNet101 method achieves only 23.60% in mIoU on the accidental scene segmentation. The prior state-of-the-art Trans4Trans model [281], employing a vision transformer and multi-source training, achieves 39.20% in mIoU. In contrast, our proposed MMUDA model surpasses all previous methods, achieving a higher mIoU of 46.97%, which is >7.50% higher than the previous state-of-the-art method. Our approach also outperforms in per-class IoU, obtaining the highest scores in 16 out of 19 categories. The improvements over Trans4Trans are particularly pronounced (>10.00% performance gain) for categories crucial to accident scene understanding, including *road*, *sidewalk*, *rider*, *car*, *truck*, and *motorcycle*. The significant improvement shows the effectiveness of the proposed MMUDA method on the DADA-seg dataset, yielding a promising solution for generalizable scene understanding.

The ablation results of the proposed modules are shown in Table 31. All experiments are based on all five multi-origin source datasets. Our baseline model with ResNet101 uses only normal source-supervised learning by aggregating multiple source domains and achieves a mIoU of 40.76%. The model with meta-learning (+Meta) further improves the mIoU by 4.26%. In addition, our proposed MDMS and transformer model with HybridASPP further improve the mIoU to 46.11% and 46.97%, respectively. These results further show the effectiveness of the proposed module in MMUDA.

6.3 CHAPTER CONCLUSION

Generalizable scene understanding aims to adapt segmentation models to real-world applications and improve their ability to handle corner cases. As the second research theme in the field of Mobility Assistance Systems (MAS), we explore novel methods to perform segmentation of transparent objects and adverse driving scenes. Transparent objects, such as glass doors and windows, are widely found in modern buildings, which are corner cases of scene recognition in practical applications. The segmentation model should be able to generalize to these unusual but mobility-relevant situations. Additionally, considering the potential synergy of helping both pedestrians and drivers, we adapt segmentation models from the perspective of walking scenes to driving scenes, including adverse and accidental cases. In this chapter, we focus on two corner-case yet safety-critical tasks:

- **Transparent object segmentation**: Segmenting partially or fully transparent objects, such as glass doors and windows, is a challenging task in the field of mobility assistance systems. These objects can blend in with the background, making them difficult to distinguish for sighted and blind people. This can pose a safety hazard, as people may not be aware of these objects and could collide with them.
- **Adverse scene segmentation**: A task of segmenting adverse driving cases, such as scenes with extreme lighting conditions, motion blur, and even traffic accident cases. These real-world conditions can make it difficult for a pre-trained model to accurately segment objects, and the dynamic and changing nature of driving scenes can further complicate the scene understanding task.

Here is a detailed overview of the contributions of each section in this chapter:

Contribution 1: We propose an efficient semantic segmentation architecture called *Transformer for Transparency (Trans4Trans)*, which uses a transformer-based encoder and decoder to unify general object and challenging transparent object segmentation in a dual-head manner. A *Transformer Parsing Module (TPM)* is proposed to fuse multi-scale representations. The proposed Trans4Trans methods achieve state-of-the-art performance on the transparent object segmentation benchmark.

Contribution 2: We advocate addressing driving scene segmentation from an adaptive perspective that jointly considers *normal*, *adverse*, and *accidental* scenarios. Trans4Trans is verified on driving scene segmentation benchmarks including Cityscapes, ACDC, and DADA-seg. Besides, we propose a novel *Multi-source Meta-learning UDA (MMUDA)* framework to perform better adaptation from multi-source domains of normal driving scenes to the target domain of abnormal accident scenes.

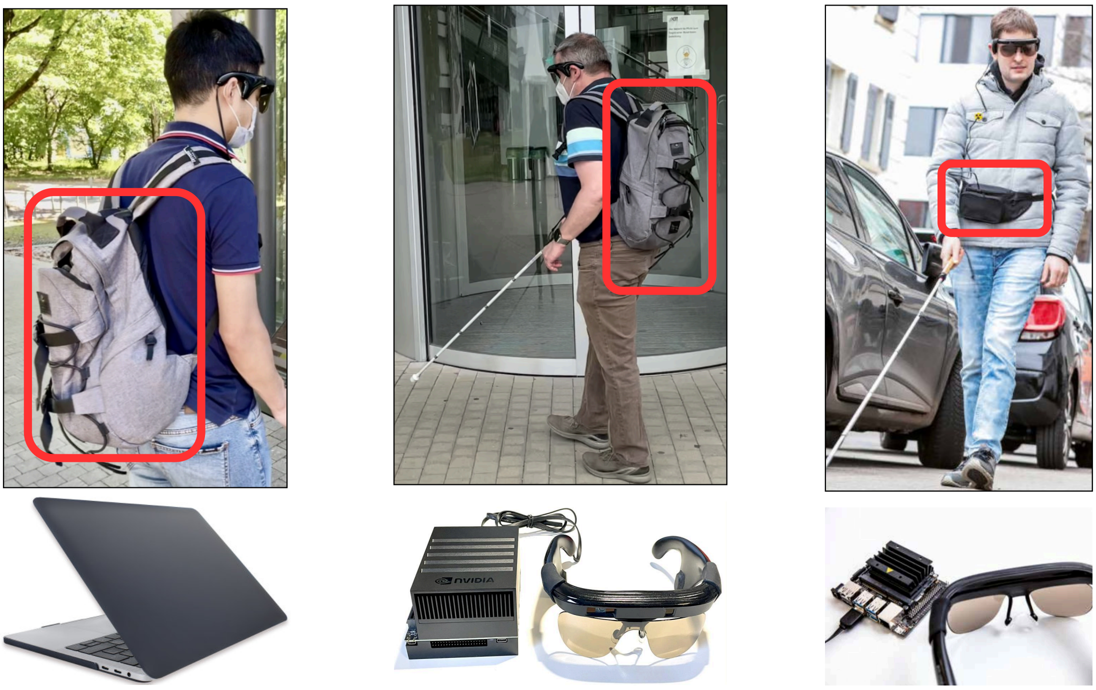
7 ASSISTIVE SYSTEMS AND APPLICATIONS

In this chapter, we focus on iteratively developing mobility assistance systems and further exploring proof-of-concept applications. We also present respective evaluations, including quantitative results, user studies, and field tests. As the third research theme in the field of MAS, one system and one proof-of-concept prototype are included: (1) We construct a human-friendly wearable system called *Vision4Blind* using semantic segmentation models. The system aims to help People with Visual Impairments (PVI) better understand their surroundings. Three versions of the system have been iteratively developed and enhanced based on feedback from the target group. The *Vision4Blind* system is detailed in Section 7.1, based on our work published in *ICCV Workshop on Assistive Computer Vision and Robotics (ACVR) 2021* [281] and in *Transactions on ITS 2022* [282]. (2) A prototype of "flying guide dog" is created by using a drone to assist PVI navigate outdoor scenes, which is presented in Section 7.2, based on our work published in *IEEE ROBIO 2021* [201].

7.1 VISION4BLIND SYSTEM

We develop the *Vision4Blind* system, a wearable assistive tool aimed at empowering People with Visual Impairments (PVI) to better understand their environment and navigate with enhanced safety and independence. According to suggestions from experts and the target group, the development of system has involved a series of iterative advancements, resulting in the creation of three versions, each surpassing its predecessor in various aspects such as functionality, portability, and usability.

As shown in Figure 58a, the first version of the *Vision4Blind* system was a heavy and bulky laptop computer that must be worn in a backpack. The laptop was powered by an embedded battery that had a very short lifespan, and the system was not very portable for the user. The laptop-based system has one advantage is to reduce the deployment effort for developers when upgrading to new models. Nonetheless, this lack of portability significantly reduces the motivation and frequency with which the target users are inclined to incorporate it into their daily lives. To enhance portability, significant efforts were invested in the exploration of suitable platforms and processors. As depicted in Figure 58b, the second iteration of the system integrated a portable processor, the NVIDIA Jetson AGX Xavier, requiring housing in a backpack due to its weight exceeding 1.5kg. While this version exhibited improved power and extended battery life,



(a) Version 1

(b) Version 2

(c) Version 3

Figure 58: Three different versions and iterative development of the *Vision4Blind* system. The versions from left to right are: (a) using a heavy laptop in a backpack, (b) using an NVIDIA Xavier Jetson AGX Xavier processor in a backpack, and (c) using a lightweight NVIDIA Jetson Nano processor in a portable belt bag with an on-device user interface.

its level of portability remained less than ideal. Figure 58c [1] illustrates the latest version of the Vision4Blind system, featuring a lightweight Nvidia Jetson Nano processor weighing only 0.25kg, comfortably accommodated within a convenient belt bag. This design significantly enhances the portability, ensuring ease of use. Additionally, the system introduces a new on-device user interface, designed for enhanced accessibility for PVI. This remarkable upgrade in portability represents a significant improvement over the previous two versions, which relied on bulky laptop computers or weighty processors. Consequently, the newest system is more user-friendly, allowing for greater convenience in both transport and use.

7.1.1 *Portable Hardware Components*

Our entire portable system consists of two hardware components: a pair of smart vision glasses and a portable GPU, *e.g.*, NVIDIA Jetson Nano.

The vision glasses (Figure 59) have been integrated with a RealSense R200 [96] RGB-D sensor to enable real-time acquisition of RGB and depth images at the resolution of 640×480, and a pair of bone-conduction earphones for delivering acoustic feedback to people with visual impairments. This is crucial as visually impaired people often rely on the sounds from the surroundings for determining the orientation and bone-conduction headphones will not block their ears when using the assistive system. The

[1] Two photographs @ Andrea Fabry are from the lookKIT magazine 2022/1: Nachhaltig digital.

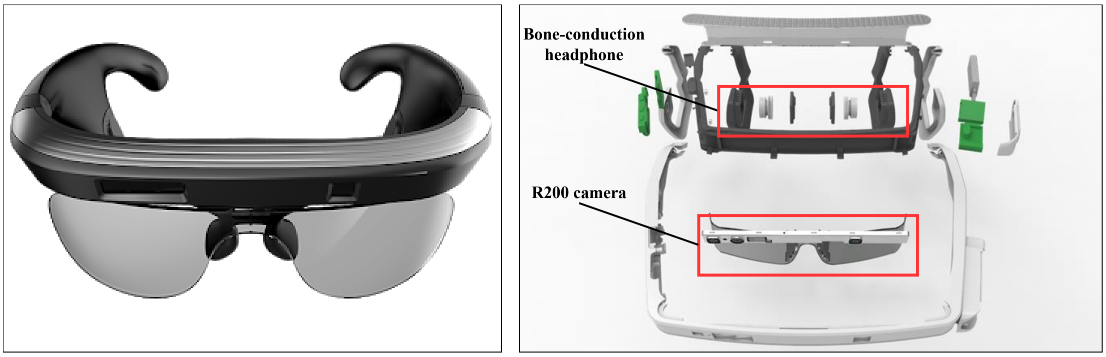


Figure 59: Detailed illustration of the glasses in our *Vision4Blind* system. The main components include RealSense R200 camera and bone-conduction headphones.

acoustic feedback primarily consists of speech, providing the name of the detected object or suggesting the orientation, such as "*glass door*" or "*left*". In texture-less indoor scenes, the projected infrared speckles (Figure 59) will augment the environments, which are beneficial for stereo matching algorithms (*e.g.*, R200 leverages a straightforward correlation engine [96]) to yield dense depth estimation. In our assistive system, depth information is mainly used to assist the obstacle avoidance function, *e.g.*, to prioritize near-range objects over mid- and long-range objects.

7.1.2 *Mobility Assistance Algorithm*

Our software components are the Trans4Trans model and a user interface as described in Algorithm 1. Starting from the input data and to guarantee the timely capture of the facing environment, the frame rate of RGB-D stream is set to 60. Once the system starts, it repeats image segmentation every n seconds. According to our experiments, the time interval setting as 2 seconds can effectively prevent cognitive overload, especially in cases of complex scenes containing many objects. Still, it is adjustable depending on the need of users, *e.g.*, a short interval for more feedback to explore unknown space.

Obstacle avoidance. When moving in a relatively restricted indoor space, the building materials or densely-arranged objects will impede the flexibility of merely using white cane as the aid tool for avoiding obstacles. In order to tackle the collision issue and balance indoor and outdoor scenarios, our system presets the highest priority for obstacle avoidance. In other words, if the average value of the depth information is smaller than the preset distance threshold $\theta_{obstacle}$, the user will be immediately notified in the form of *vibration*. To minimize the uncertainty of vibrations and the cognitive load, only one single default threshold is set to 1 meter, instead of setting various vibration frequencies for different distances. Another purpose is to preclude the chaotic and low-confidence segmentation from the less-textured images when users walk too close and face to the object surface, such as images from white wall or doors.

(Transparent) object segmentation. After receiving the RGB image $X \in \mathfrak{R}^{H \times W \times 3}$, our efficient Trans4Trans model outputs two segmentation predictions, which are general object segmentation $G \in \mathfrak{R}^{H \times W \times 13}$ and transparent object segmentation $T \in \mathfrak{R}^{H \times W \times 11}$, respectively. The general object segmentation is divided into G_{path} for *walkable path* and G_{object} for other *object* classes. For example, the system can out-

Algorithm 1 : Assistive system

```
Data : RGB-D as $X \in \mathcal{R}^{H\times W\times 3}$ and $Y \in \mathcal{R}^{H\times W}$.
Result : General segmentation $G \in \mathcal{R}^{H\times W\times 13}$; Transparency $T \in \mathcal{R}^{H\times W\times 11}$;
initialize walkable rate: $R_l, R_f, R_r$, parameters: $\theta_{obstacle}, \theta_{trans}, \theta_{walkable}$ ;
while system start and each $n$ seconds do
    RGB-D update and Trans4Trans segmentation:
    $G_{path} \in \mathcal{R}^{H\times W}, G_{object} \in \mathcal{R}^{H\times W\times 12}$ ;
    $T_{stuff} \in \mathcal{R}^{H\times W\times 3}, T_{thing} \in \mathcal{R}^{H\times W\times 8}$ ;
    partition $\{R_l, R_f, R_r\} \leftarrow G_{path}$ ;
    if $\overline{Y} < \theta_{obstacle}$ then
        $vibration$ as obstacle warning;
    else if $max\{\overline{T}_i\} \in T_{stuff} > \theta_{trans}$ then
        $speech \leftarrow argmax\{\overline{T}_i\} \in T_{stuff}$ ;
    else if $max\{R_l, R_f, R_r\} > \theta_{walkable}$ then
        $speech \leftarrow argmax\{R_l, R_f, R_r\} \in \{left, forward, right\}$;
    else
        $speech \leftarrow nearest\{T_{thing}, G_{object}\}$;
    end
end
```

put a speech of "*left*" or "*glass door*". Afterwards, the walkable mask is further partitioned into three regions as $\{left, forward, right\}$ directions for orientation. In order to correct the wrongly-segmented walkable area by the high-confidence transparency perception, the transparent object segmentation is divided into two disjoint sets as: $T_{stuff} \in \mathcal{R}^{H\times W\times 3}$ with {*window, glass door, glass wall*}, and $T_{things} \in \mathcal{R}^{H\times W\times 8}$ with {*shelf, jar/tank, freezer, eyeglass, cup, bowl, bottle, box*}.

Walkable path detection. After achieving object segmentation, the local ratio of walkable area G_{path}, *e.g.*, *floor* category from Stanford2D3D, is further horizontally divided into three different directions as $\{R_l, R_f, R_r\} \leftarrow G_{path}$. Then, an intuitive and effective strategy is to prompt the direction that has the largest walkable area, only when its local ratio is greater than the preset threshold $\theta_{walkable}$ for safety. The output is one speech of {"*left*", "*forward*", "*right*"}. According to our test, this orientation approach guarantees anti-veering in a straight path outdoors and indoors. Furthermore, it can also accurately predict the best instantaneous turning direction during walking at an intersection, so as to constantly yield a safer direction suggestion.

7.1.3 *User Study and Field Test*

Based on our publications in *ICCV Workshop on Assistive Computer Vision and Robotics (ACVR) 2021* [281] and in *Transactions on ITS 2022* [282], we have two rounds of qualitative study with 5 participants and 3 experts to assess the acceptance of our *Vision4Blind* prototype and draw design conclusions [143].

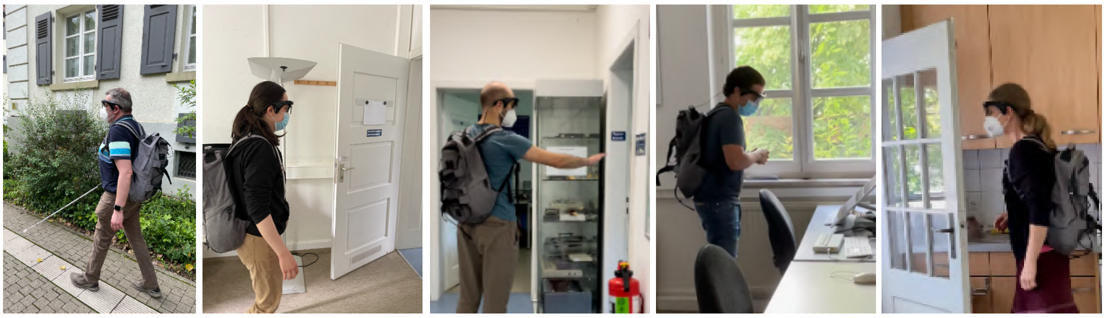


Figure 60: Incidences of participants using the system for navigation outdoors and indoors.

Gender		Age Range				Hearing loss
Male	Female	18-25	26-35	36-45	46-55	No
3	2	1	2	1	1	5

Table 32: Aggregated demographics of participants.

7.1.3.1 *Methodology of Study*

As mentioned in Algorithm 1, the *Vision4Blind* system based on the version of Trans4Trans model has three main functions partitioned into four steps as mutually exclusive: (1) The obstacle avoidance based on depth information (*i.e.*, <1.0 m) has the highest priority. (2) Three different types of transparent stuff (*wall*, *door*, and *window*) will be alerted via speeches. (3) Walkable path will be indicated in three different directions (*left*, *forward*, and *right*). (4) Other general objects and transparent things will be fed back. Participants tried the system inside 2 buildings, and the blind participant also on a 700m route outdoors – see Figure 60. The study lasted about 2 hours. As Corona-protective measures, everyone wore FFP2 or surgery masks throughout the study and the prototype was disinfected several times. After a short introduction (Figure 61) of the system and it functionality, all participants agreed to participation and recordings of the whole study and they signed the data protection statement respectively. First, the participants put on the *Vision4Blind* system. They were allowed to seek assistance while wearing the device. Then, they walked around the rooms, thinking out loud [91]. The study was recorded with an action camera and voice recorder. At the end, demographics and NASA Raw Task Load Index (RTLX) [72] questionnaires were filled in.

7.1.3.2 *Demographic of Participants*

In the first user study, we evaluated the system with 5 participants [281], one of whom was an expert, and another one was expert and blind user at the same time. Age and gender of five participants are in Table 32. We subsequently repeated the experiment with 3 further sighted experts, and we only report here the aggregated results from the 5 experts: E1B (early blind expert), E5-E8 (sighted experts). When asked if they can see glass objects, E1B said he can sometimes see some light-dark contrasts, which allows him to perceive closed windows. Windows that open inside the room, however, are very dangerous, according to E1B, as one can get serious head injuries. All sighted participants said they can see glass objects, but some of them, like glass doors, glass walls or windows, can be challenging under particular conditions (E5).

Figure 61: Introduction and tutorial session on the use of the Vision4Blind system.

7.1.3.3 *Cognitive Load*

The RTLX, averaged over the five expert participants, was 16.3 with a standard deviation of 8.1. The range is from 0 to 100, the lower the better. This score is enough to keep the user motivated, while not burdening too much [139]. This score, however, must be critically interpreted, since it might not be representative for the users wearing the system in their daily activities. Instead, this score might reflect the cognitive load of the experts assessing the system, since this was their task, and not simulating user behavior. Only the score of the blind participant is highly relevant for the cognitive load of users wearing the system. This score is 13.3, thus very close to the average, but being alone, it has hardly any statistical relevance. More studies will have to be performed in the future to assess the cognitive load of the users wearing the system. According to the individual ratings, effort and physical demand were slightly higher, while frustration was the lowest subscale. This might suggest that users enjoyed the experience of using our system, but a further reduction of hardware would be welcome.

7.1.3.4 *User Comments and Qualitative Analysis*

A thematic analysis [13] performed on the comments made by the experts (both recorded and from the questionnaires) yielded the following insights:
Functionality. All experts found the system useful and were impressed by its functionality, for instance,

> *"For the first time, I had the feeling that artificial intelligence can be useful [...]. [I liked] how much it recognized correctly. [...] Systems react much better [than 10 years ago...]. I think it's just cool!"* — (E1B)

> *"I'm very excited about how it works, it says a lot of things [...] It's yeah, I'm really really impressed [...] and works really smoothly. [I like] that the 3 functions are so smoothly integrated."* — (E5)

> *"But great system, yes, great the information [that one gets]."* — (E6)

> *"It was a surprise for me to detect also laptop and this kind of stuff. The CV detection is wonderful. [...] List of objects relatively broad, [it was] a surprise. Very interesting, exactly what [a blind person is] interested in: person, chair - the first thing I would look for in a new room during a meeting."* — (E7)

> *"I think it's cool, for blind [people] it's cool."* — (E8)

We found that most positive comments are on the amount and type of objects recognized by the *Vision4Blind* system (E1B, E5, E7, E8).

When asked "*what did you think about the system?*", E8 mentioned the object recognition. Two experts mentioned further objects that they consider very important and should be included in future implementations, namely: *trash cans* outdoors and *city scooters* (E1B) and *construction site fences* (E5). Expert E6 thought that *low obstacles*, within 1 meter from the user, and which are not covered by the camera glasses range, should also be detected: *"These low obstacles are dangerous"* (E6). E6 proposed to have a second camera mounted on the abdomen area. E5 and E8 suggested to have a camera mounted on a white cane. This can reduce the amount of hardware that a user carries.

The experts gave some important suggestions on subsequent system development, such as identifying more objects (E1B, E5), mounting a second camera to detect low-lying obstacles (E6), and hinting the directions of detected objects (E1B, E5-E7). Two experts (E5, E7) commented positively upon the free path detection, and mentioned that the obstacle detection should be improved. Most issues with the obstacle detection came from the 2 seconds cycles (frame aggregation), which often caused a delay and delay inconsistencies (E1B, E5, E6, E7). To tackle this problem, it is desirable to further decrease the system response time. Regarding the suggestions from E5 and E8, adaptive feedback cycles for different functions can be implemented. For example, the feedback of obstacle detection should be given generally faster than for the other two functions. The default in this case could be for instance 1 second instead of 2. Besides, the distances of detected objects are helpful for keeping social distances in COVID-19 pandemic times (E6, E8). The comments on the *obstacle detection* were divided.

> *"For me, if it should be my assistive technology, I could buy it if the obstacle would be improved a little bit. Because the detection is nice and I would like to have this device, but the distance of object needs a little improving."* — (E7)

Expert E5, on the other hand, thought that the obstacle detection could also be useful for keeping the distance in COVID-19 pandemic times:

> *"I think you can also use it for social distancing, because it says when someone is in front of you, so you know ok, there's a person."* — (E5)

Three experts (E5-E7) considered that this system is a nice complement to the white cane, but should not be used as alternative or hard to replace the white cane with it.

> *"Below you have the white cane, at the distance that is not covered by the [system's] camera. They complement each other well."* — (E5)

> *"When the accident of me, was [...] that was in the upper area, that's why this [system] is good, because it already covers the upper area, and below is secured by the white cane so to speak."* — (E6)

> *"One must have a white cane; because only with the system, a blind person wouldn't feel so safe."* — (E7)

Hardware. The hardware was perceived as quite light weight (E5, E6), and in any case much better than previous prototypes (E1B) tried out by the experts in the past (at least three out of five had tried similar prototypes in the past). However, two experts (E1B, E5) considered the hardware still too big for a real-world deployment:

> *"[use] a belt instead of a backpack [and] Bluetooth instead of cables for the glasses. [...] The glasses look good."* — (E1B)

> *"Ideally, it should run on a phone."* — (E5)

> *"Wearing the camera as a pair of glasses is very comfortable, even though it is thick."* — (E1B)

Besides, Experts E5 and E6 also commented positively on the system's battery life, which can last for an entire day.

Interface. Four out of five experts thought the interface was very intuitive. Only E8 was neutral with respect to this.

> *"The object announcement. The acoustic signal is easy to follow + easy to understand."* — (E6)

> *"The synthetic voice was very helpful, because it differentiates well from background noise."* — (E1B)

Context of use. Expert E7 thought the system is good for getting an overview of a new room, but not so good for known rooms. He also suggested implementing objects searching and counting. Both E5 and E7 thought the system can be used for social distancing, but referred to two different functions of the system, namely obstacle detection and free path recognition. E5 suggested to use the system also for sighted people for warning when walking while looking at the phone.

Control. E7 thought the user should be in full control of the system, like it is the case with the white cane: *"I can mute it when I can't interact with the system in certain situations - the white cane does what I want"*. Both E5 and E8 thought it is important to have the option to turn functions on and off, or switch to different modes (E8).

7.1.3.5 *User Study Conclusion*

The functionality offered by the *Vision4Blind* system so far can be of great use to people with visual impairments. All experts were positive about the system. Especially the object recognition was appreciated. Some improvements were suggested, such as conveying to the user the distance and direction of objects, covering more objects, and improving the efficiency for obstacle avoidance, etc. Also important, the users should be able to configure the system as much as possible and turn functions on and off as they need, or change the way things are conveyed (*e.g.*, speech, sonification, vibration). Due to the heterogeneity of the user group, the configurability is a very important aspect. Based on the comments and the qualitative analysis, all valuable suggestions and novel ideas will be carefully considered in the development of the next generation of mobility assistance system. We are committed to developing a system that is both user-friendly and effective, and we appreciate the feedback from our users and participants.

7.2 FLYING GUIDE DOG PROTOTYPE

This section is based on our work published in *IEEE International Conference on Robotics and Biomimetics (ROBIO) 2021* [201]. The included methodological and technical contributions result from collaboration with a co-supervised practical course project.

7.2.1 *Drone Assistance System*

To develop a spatially flexible mobility assistance system, we build a novel "flying guide dog" prototype (Figure 62), exploring the combination of drone and scene understanding. According to previous work [6], drone navigation is more accurate and faster as it gives a continuous and physical feedback in the direction of travel. Performing semantic segmentation on frames captured by the drone camera, a variety of ambient visual information can be extracted, such as *sidewalks*, *crosswalks*, and *traffic lights*. Based on its perception of the environment, the drone adjusts itself and leads the user to walk around safely. To follow the drone, the user holds a soft string attached to the drone. Besides, the user receives voice prompts via a Bluetooth bone conduction headphone.

Discovering the walkable path is one of the major functions of our prototype. Based on the segmentation prediction, walkable area (*e.g. sidewalks*, *crosswalks*) can be discovered by their corresponding colors. In order to make the drone keep flying along the walkable path safely, we develop a control algorithm so that the drone can automatically adjust its direction and velocity according to the estimated centroid of the sidewalk. Another function is assisting the user to pass the pedestrian traffic light, *i.e.* street crossing. Our prototype not only distinguishes pedestrian crossing lights from other types of traffic lights but also recognizes their colors. Since there is currently no dedicated traffic light dataset containing both pedestrian and vehicle traffic lights, we introduce a new dataset called *Pedestrian and Vehicle Traffic Lights*. To verify the effectiveness, we further conduct a user study in real-world scenarios. The result indicates that our prototype is effective for visually impaired assistance and easy to use.

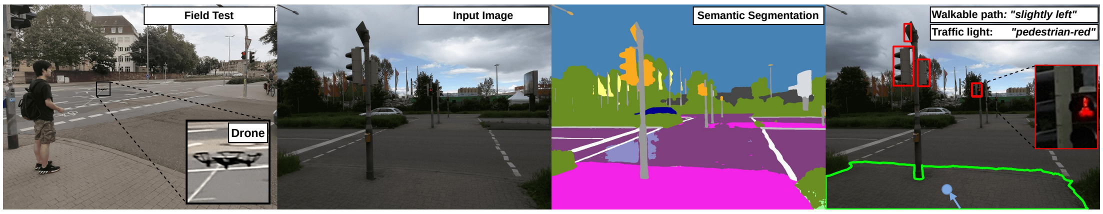

Field Test
Drone
Input Image
Semantic Segmentation
Walkable path: "slightly left"
Traffic light: "pedestrian-red"

Figure 62: The "flying guide dog" prototype. Images from left to right are: field test in front of an intersection, input image from the drone's perspective, semantic segmentation result, interpreted result with centroid, boundary and bounding boxes for the drone control algorithm.

7.2.2 *Hardware Design and Algorithm*

The drone is connected to the computer using DJITelloPy[1] library via Wifi. For each frame captured by the drone's camera, the segmentation model outputs a prediction. To make the drone fly along the walkable path, the largest walkable area is extracted and

[1] DJITelloPy: `https://github.com/damiafuentes/DJITelloPy`.

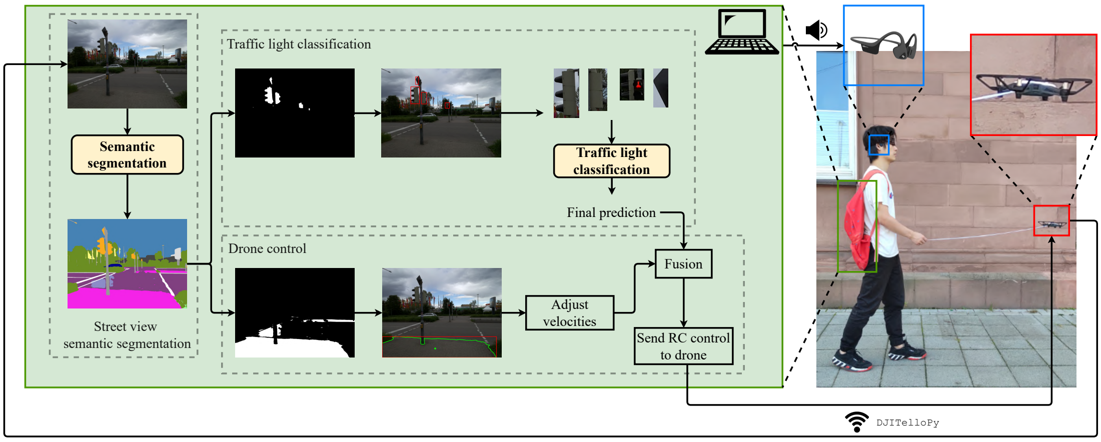


Figure 63: The "flying guide dog" system overview. There are three components: street view semantic segmentation, traffic light classification, and drone control.

its centroid is then estimated. On the basis of centroid estimation, velocity adjustment is computed. Meanwhile, if traffic light is detected, we crop them out and input into the classification model. Fusing the classification prediction and velocity adjustment, a remote control command is sent to the drone. Additionally, the user gets voice prompts via a Bluetooth-connected bone conduction headphone. As illustrated in Figure 63, our system has three modules: (1) semantic segmentation, (2) traffic light classification, and (3) drone control. These modules will be detailed below.

7.2.2.1 *Semantic Segmentation Model*

A vital requirement for the segmentation model is real-time performance. According to the comparison of real-time semantic segmentation models, SegFormer-B0 [238] has significant advantages on speed and accuracy. Furthermore, it is more robust to common corruptions and perturbations. While Cityscapes [44] are recorded in a unified setting, Mapillary Vistas [142] are globally taken by diverse devices from different viewpoints. Moreover, comprising 25,000 densely annotated street level images into 66 categories, Mapillary Vistas is 5× larger than Cityscapes in terms of fine-grained annotations. Therefore, Mapillary Vistas is more diverse and promising for yielding robust models, thus suitable for our prototype.

7.2.2.2 *Traffic Light Classification*

After semantic segmentation, a prediction mask regarding traffic lights can be obtained. This mask can then be used to crop a traffic light patch from its full-resolution image. Image classification is more computationally efficient in a cropped patch than in a full-scale image. To simplify and unify the recognition of the color and category of traffic lights, we train a light CNN. This CNN is expected to be more accurate and efficient, and it should also eliminate the influence of vehicle traffic lights. However, there is currently no dataset that is directly suitable for our task. To facilitate traffic light classification, we introduce a new dataset called *Pedestrian and Vehicle Traffic Lights* with the goal of distinguishing pedestrian traffic lights from vehicle traffic lights. We perform four steps to collect this dataset. Firstly, we crop the traffic lights from Cityscapes

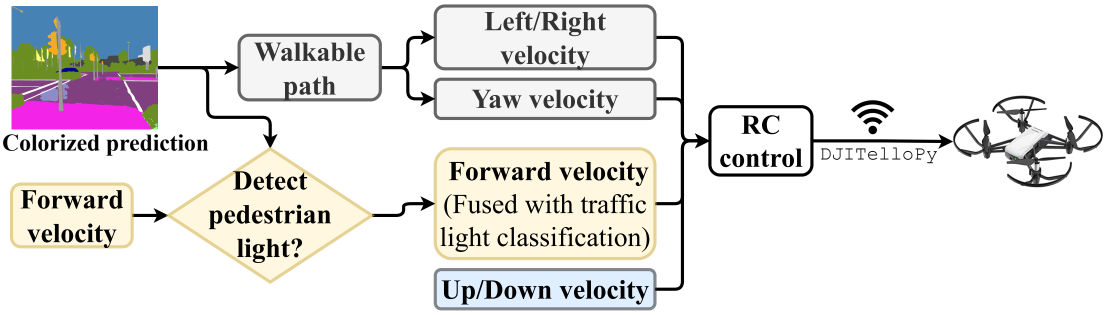


Figure 64: Drone control overview. Four types of velocity are updated based on the walkable path and the traffic light prediction.

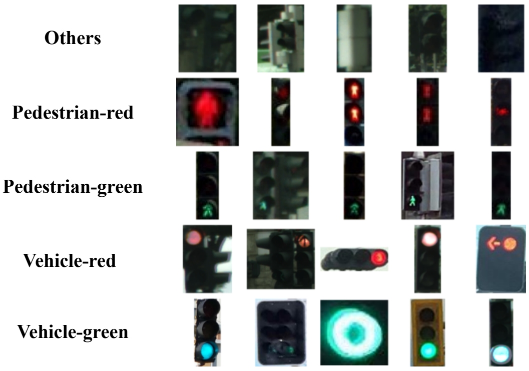


Figure 65: PVTL dataset with 5 categories.

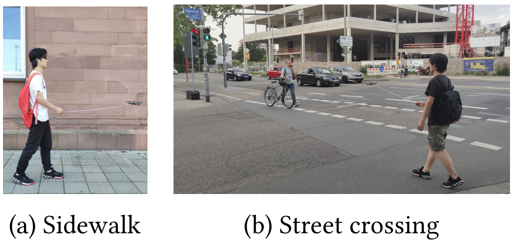

(a) Sidewalk (b) Street crossing

Figure 66: The user study consists of two scenarios: sidewalk and street crossing. As a test for the proof of concept, the sighted participant wears a blindfold during the user study.

[44], Mapillary Vistas [142], and Pedestrian lights [167] based on their annotations. Secondly, we clean up those image patches with resolution smaller than 8×8. Then, we manually annotate the images into 5 categories: *pedestrian-red*, *pedestrian-green*, *vehicle-red*, *vehicle-green*, *others*. Last but not least, a class balancing is performed to maintain 300 images for each class. We visualize some representative examples in Figure 65. The category *others* refers to the side and back of the traffic lights or traffic lights that are not illuminated. All data and annotations will be made publicly available.

7.2.2.3 *Drone Control*

After achieving semantic segmentation and traffic light classification, four types of velocity are calculated to control the drone, as depicted in Figure 64. Among them, the up/down velocity v_{ud} is obtained by the Tello's vision positioning system to maintain the flying height h_{target}, which is preset as 1.2m to ease the user interaction. Finally, RC commands are sent to the drone using DJITelloPy. A detailed description of the control strategy is presented in Appendix A.2.

7.2.3 *User Study and Discussion*

To evaluate the assistance functions of our system, an user study in real world scenarios is conducted using NASA Task Load Index (NASA-TLX) method [72].

Aspects	Statements
Orientation	I can walk in the proper direction.
Position	I can walk in the middle of the path.
Traffic Light	I can cross the road at the right time according to the pedestrian lights.
Learnability	I can get familiar with this system easily.
Mentally easy to use	I don't need much time to think to follow the drone.
Physically easy to use	I don't need to put in a lot of physical effort when using the system.

Table 33: Six aspects are included in the questionnaire.

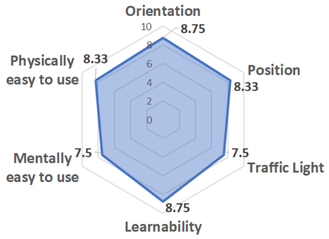


Figure 67: User study evaluation.

7.2.3.1 *User Study Setup*

During the user study, the laptop is placed in a backpack. Six participants aged between 24 and 35, including 5 males and 1 female, are sighted but blindfolded during the test. The drone flies in front of the participants to simulate the real guide dog and to get an unobstructed view of the environment. The participant follows the drone by feeling the traction from a string attached to the drone. Our prototype is evaluated in two open-space scenarios. The first scenario focuses on walking along the *sidewalk* (Figure 66a) in 20 meters length. Some obstacles, such as *bicycles* and *pedestrians*, were randomly appeared to test the obstacle avoiding function. The second scenario, an intersection (Figure 66b) including vehicle and pedestrian traffic lights, is selected to test the street crossing functionality according to the pedestrian light prediction. After testing, participants fill questionnaires as in Table 33. They score each aspect from 1 to 10, among which 1 and 10 mean *strongly disagree* and *strongly agree*, respectively.

7.2.3.2 *Evaluation*

A radar graph assessing multiple aspects is shown in Figure 67. For features of walking along the walkable path, *Orientation* and *Position*, participants find it reliable for guiding and obstacles avoidance. In terms of *Physically easy to use* and *Learnability*, they make positive comments. After a brief introduction, the participants are able to use the system alone. However, the *Traffic light* feature still needs to be improved. It could be harder to cross the street with multiple pedestrian lights. To tackle this issue, the speed of drone can be adjusted during crossing via the drone control algorithm. The participants also rate the feature *Mentally easy to use* lower than others, as sometimes it is insufficient to feel the traction from the connected string. This issue is subsequently improved with additional voice feedback using a wireless bone conduction headphone, so as to alert users the walking direction and the color of the pedestrian light. Despite a limited number of participants, most of them comment that the prototype is helpful, which provides a hint about the effectiveness of our "flying guide dog" concept.

This proof-of-concept prototype was implemented to assess the potential of using new techniques to create assistive systems. However, the target group of people with visual impairments was not included in the user study at the concept proof stage. Additionally, the prototype is limited by the drone's battery capacity, which only supports a maximum flight time of 13 minutes. Another major limitation is that the drone is too light to resist wind. One reasonable approach to address these issues would be to use a more powerful drone with a larger battery capacity.

7.3 CHAPTER CONCLUSION

Assistive systems and applications have being developed to provide scene understanding and navigation support for People with Visual Impairments (PVI). However, the challenge arises as scene understanding models typically demand powerful but heavy hardware platforms, making the design of a portable and user-friendly system quite challenging. As the third research theme in Mobility Assistance Systems (MAS), we spent large effort to an iterative development process for three versions of the *Vision4Blind* system. Furthermore, in pursuit of innovative assistance solutions, we look into the possibility of using Unmanned Aerial Vehicle (UAV) systems, presenting a conceptual prototype as our initial trial. To verify these systems and applications, our evaluation includes a series of quantitative results, qualitative insights from user studies and field tests. This chapter focuses on two different systems.

- **Wearable assistance systems**: In the process of deploying semantic segmentation models to real-world scenarios, the development of a portable and user-friendly prototype system is crucial. Combined with scene understanding models and portable hardware devices, the wearable system can provide mobility-related information to assist people with visual impairments to navigate more independently and safely.
- **Flying guide dog**: Utilizing UAVs or drones as tools to assist people with visual impairments in navigation represents a new concept in assistive technology. UAVs can have more flexible relative poses than wearable systems, and can provide perception assistance in front, above, below, or even around the user. Besides, it can provide more intuitive physical interaction through a wired connection or provide a hybrid communication via a wireless connection.

Here is a more detailed summary of the contributions of each section in this chapter:

Contribution 1: We iteratively develop and test a wearable assistive system with a pair of smart vision glasses and a portable GPU, based on our proposed vision transformer models. This system, called *Vision4Blind*, has been validated with a diverse user group. The results of the user study show that it is easy to use and can provide accurate scene understanding information to assist people with visual impairments. Furthermore, qualitative analysis results and comments from target users are summarized for the iterative development of the *Vision4Blind* system.

Contribution 2: We propose a novel “flying guide dog” prototype that uses a drone and semantic segmentation to help people with visual impairments navigate their surroundings. The drone is equipped with a camera and a model that can identify walkable paths and traffic lights. A control algorithm is proposed to enable the drone to fly along the walkable path automatically and to interact with the user via voice feedback. To improve the recognition of traffic lights, a new dataset is created to train a model for fine-grained traffic light classification.

Part IV

INSIGHTS OF SCENE UNDERSTANDING

8 CONTRIBUTIONS

> This thesis delves into the vision of smart mobility through exploration in two research domains: *intelligent transportation systems (ITS)* and *mobility assistance systems (MAS)*. In the field of ITS, the holistic and robust scene understanding is pursued through panoramic and multimodal semantic segmentation. Meanwhile, considering the synergy of driving and walking scenarios, the navigational and realistic scene understanding in the field of MAS is propelled by research themes of visual localization, semantic mapping, transparent object segmentation, and adverse scene segmentation. As real-world applications, the *Vision4Blind* system has being iteratively designed to facilitate navigation and mobility for people with visual impairment. All contributions of this thesis have been made publicly available for the research community.

8.1 NEW DATASETS

DensePASS dataset for panoramic semantic segmentation [280]. DensePASS is a new dataset created for advancing panoramic semantic segmentation towards holistic scene understanding for intelligent vehicles. For the first time, the dataset includes 100 labelled images for testing panoramic semantic segmentation and 2000 unlabelled images for training in the domain adaptation manner. It addresses the lack of established benchmarks for the challenging task of Pinhole-to-Panoramic recognition. The images are captured using Google Street View and include scenes from different continents. The data is manually annotated with 19 categories that are also present in the pinhole camera dataset Cityscapes and other prominent semantic segmentation benchmarks. Overall, DensePASS stands as a significant and valuable addition to the domain of panoramic segmentation for driving scenes. This dataset not only offers a substantial collection of labelled and unlabelled images but also embodies a promising design that directly tackles the intricacies of Pinhole-to-Panoramic Unsupervised Domain Adaptation (UDA). As such, we believe that DensePASS will emerge as an indispensable asset for researchers and developers actively engaged in addressing this critical challenge. The dataset has been made publicly accessible at the following URL: https://github.com/chma1024/DensePASS.

DeLiVER dataset for multimodal semantic segmentation [279]. The DeLiVER dataset is created specifically for the task of arbitrary-modal semantic segmentation. Based on the CARLA simulator, the dataset incorporates a diverse range of data sources including Depth, LiDAR, Views, Events, and RGB images. Besides, the dataset also encompasses scenarios involving five distinct sensor failure cases, comprising Motion

Blur (MB), Over-Exposure (OE), Under-Exposure (UE), LiDAR-Jitter (LJ), and Event Low-resolution (EL). These cases serve to validate the robustness and stability of model performance in the face of sensor malfunctions. The sensors are positioned at various locations on the ego car to provide multiple viewing angles, such as *front*, *rear*, *left*, *right*, *up*, and *down*. Each individual sample within the dataset is annotated with both semantic and instance labels. A total of 25 distinct classes are included in DeLiVER dataset: *Building, Fence, Other, Pedestrian, Pole, RoadLine, Road, SideWalk, Vegetation, Cars, Wall, TrafficSign, Sky, Ground, Bridge, RailTrack, GroundRail, TrafficLight, Static, Dynamic, Water, Terrain, TwoWheeler, Bus, Truck*. With its comprehensive viewpoints, diverse scenarios, multiple modalities, and detailed annotations, we believe that the DeLiVER dataset has the potential to make a significant contribution to the task of arbitrary-modal semantic segmentation, enhancing the robust scene understanding. The dataset has been made publicly accessible at the following URL: https://jamycheung.github.io/DELIVER.html.

8.2 NEW METHODS

P2PDA Model [280]. Towards *holistic* scene understanding via *panoramic* semantic segmentation, we propose P2PDA, an innovative framework designed for achieving 360° perception of self-driving scenes. This is accomplished by leveraging the process of adapting semantic segmentation networks from a source domain rich in labels, comprising standard pinhole camera images, to an unlabelled target domain involving panoramic data. Our P2PDA framework has an encoder-decoder based semantic segmentation network, along with four distinct building blocks for facilitating domain alignment: the Segmentation Domain Adaptation Module (SDAM), Attentional Domain Adaptation Module (ADAM), Regional Context Domain Adaptation Module (RCDAM), and Feature Confidence Domain Adaptation Module (FCDAM). These modules are strategically positioned at two different network stages, both after and before the decoder of the segmentation network. The code and model weights have been released at the following URL: https://github.com/chma1024/DensePASS.

Trans4PASS Model [283]. To tackle distortion and deformation of panorama images, we present a novel architecture named Transformer for Panoramic Semantic Segmentation (Trans4PASS). This architecture effectively addresses image distortions and object deformations through two innovative design choices: (1) We integrate the Deformable Patch Embedding (DPE) at both the initial image sequentialization stage and the intermediate feature interpretation stage. This empowers the model to capture characteristic panoramic image distortions while preserving semantic information. (2) Within the feature parsing stage, we introduce the Deformable MLP (DMLP) module. This module enhances global context modeling by incorporating patches with learned spatial offsets, thereby improving the capacity for comprehensive context understanding. The code and model weights have been released at the following URL: https://github.com/jamycheung/Trans4PASS.

CMX Model [278]. Towards *robust* scene understanding via *multimodal* semantic segmentation, we propose CMX, a universal cross-modal fusion framework for RGB-X semantic segmentation in a novel interactive fusion manner. The RGB-X modalities can be RGB-Depth, -Thermal, -Polarization, -Event, -LiDAR data. The CMX framework is structured as a two-stream architecture comprising RGB and X-modal streams. Furthermore, the model incorporates two specific modules for the purpose of feature interaction and feature fusion between these streams. (1) The Cross-Modal Feature Rectification Module (CM-FRM) can recalibrate the bi-modal features by leveraging their spatial and channel correlations. (2) The Feature Fusion Module (FFM) is constructed in two stages and it performs sufficient information exchange before merging features. The code and model weights have been released at the following URL: https://github.com/huaaaliu/RGBX_Semantic_Segmentation.

CMNeXt Model [279]. As an advanced version of CMX model, we present the CMNeXt model for arbitrary-modal semantic segmentation (AMSS). While adding modalities, CMNeXt effectively manages the computational overhead, thanks to the Hub2Fuse paradigm. CMNeXt follows an asymmetric structure, featuring two branches: one for RGB and another for supplementary modalities. In particular, the hub step of Hub2Fuse paradigm entails a Self-Query Hub (SQ-Hub) for gathering complementary insights from auxiliary modalities. The SQ-Hub dynamically selects informative features before fusing them with the RGB branch. Another significant advantage of the SQ-Hub lies in its extensibility to accommodate any number of modalities, with only a negligible increase in parameters. Additionally, we leverage cross-fusion modules from CMX and combine them with our newly devised Parallel Pooling Mixer (PPX). These design decisions coalesce within the CMNeXt architecture, offering a pathway for Arbitrary-Modal Semantic Segmentation (AMSS). By thoughtfully integrating diverse modalities, CMNeXt is capable of mitigating individual sensor failures and enhancing the overall segmentation robustness. The code and model weights have been released at the following URL: https://jamycheung.github.io/DELIVER.html.

MatchFormer Model [228]. Towards *navigational* scene understanding, a novel method MatchFormer is proposed, which helps to achieve multi-wins in precision, efficiency, and robustness of feature matching across indoor and outdoor pose estimation and localization tasks. To improve computational efficiency and foster robust matching in low-texture scenarios, we propose interleaving self- and cross-attention mechanisms within MatchFormer, thereby constructing a matching-aware encoder. This innovative approach, referred to as extract-and-match, entails simultaneous learning of both the local features of an image itself and the similarities between its paired images. This strategic interplay alleviates the burden on the decoder, resulting in an overall streamlined model. The strategic positioning of cross-attention in the earlier stages of the encoder significantly bolsters feature matching, particularly in challenging contexts such as low-texture indoor scenarios or when dealing with fewer training samples in outdoor settings. This design choice renders MatchFormer particularly well-suited for real-world applications where collecting and annotating large-scale data is often infeasible. The code and model weights have been released at the following URL: https://github.com/jamycheung/MatchFormer.

Trans4Map Model [25]. Performing navigation and advanced scene understanding requires an accurate top-down semantic map generated from perspective views. To realize this, we reexamine top-down semantic mapping through transformer-based models and propose the novel Trans4Map framework. It delivers two primary benefits: (1) The long-range feature modeling ability is advantageous to obtain a more comprehensive spatial representation; (2) The efficient and lightweight model structure enables the one-stage end-to-end mapping pipeline. Our Trans4Map framework includes 3 steps: (1) The incoming N egocentric images are fed into the transformer backbone; (2) The Bidirectional Allocentric Memory (BAM) module projects the extracted feature; (3) A lightweight CNN-based decoder parses the projected feature and predicts the allocentric semantics. The code and model weights have been released at the following URL: `https://github.com/jamycheung/Trans4Map`.

360BEV Model [204]. To provide holistic and accessible map for advanced scene understanding, we introduce a novel 360BEV model to generate bird's-eye-view semantic mapping, *i.e.*, predicting a complete BEV semantic map from a single-frame 360° image with depth. By decoupling the computationally expensive processing of sequences or multiple views, our 360BEV semantic mapping is more streamlined for generating indoor semantic maps. To enable 360BEV segmentation we present two real indoor BEV datasets, which are extended from the Matterport3D and Stanford2D3D datasets. First, the Front-View images captured by pinhole cameras from Matterport3D are extended to 360° panoramas for benchmarking on 360FV-Matterport. Furthermore, for the first time, two BEV datasets, 360BEV-Matterport and 360BEV-Stanford are established to enable 360° bird's-eye-view semantic mapping. The data and model weights have been released at the following URL: `https://jamycheung.github.io/360BEV.html`.

Trans4Trans Model [282]. To resolve the safety-critical object recognition, we present Transformer for Transparency (Trans4Trans), an efficient semantic segmentation architecture with dual heads. Trans4Trans is established with both transformer-based encoder and decoder to fully exploit the long-range context modeling capacity of self-attention layers. In particular, Trans4Trans includes a novel Transformer Paring Module (TPM) for fusing multi-scale feature maps. The symmetric transformer-based decoder can consistently parse the feature maps from encoder. By incorporating the capability to semantically predict common object classes such as walkable areas, our system becomes adept at accurately and comprehensively segmenting both transparent objects and general objects. The code and model weights have been released at the following URL: `https://github.com/jamycheung/Trans4Trans`.

MMUDA Model [133]. To delve deeper into the abnormal scene segmentation from the traffic scene, we propose a novel *Multi-source Meta-learning UDA* framework to transform models to the unusual target scenes. Our framework learns from the label-rich datasets of conventional and normal driving scenes (*i.e.*, *source* domain), and then automatically adapts to abnormal accident scenes (*i.e.*, *target* domain) with only unlabelled training data. To effectively learn from the entire unlabelled target domain dataset, we put forward a *Multi-Domain Mixed Sampling (MDMS)* strategy, which can augment the training samples of multiple source domains. The code and model weights have been released at the following URL: `https://github.com/xinyu-laura/MMUDA`.

8.3 NEW SYSTEMS

Vision4Blind System [281]. Towards *realistic* scene understanding, we propose a wearable assistance system, *Vision4Blind*, which is capable of performing real-time wayfinding and object segmentation to assist people with visual impairments travel safely. To obtain human-friendly design and good use experience, our *Vision4Blind* wearable system comprises a pair of smart vision glasses and a portable GPU processor. Based on the proposed Trans4Trans model, the system can deliver a realistic scene understanding swiftly and accurately thanks to the high efficiency of our model. With the complete semantic information, the user interface consists of a customized set of acoustic feedback via sonification of detected objects, walkable directions and warnings of the obstacles, which yields intuitive suggestions and no prior knowledge is needed. The system implementation have been made publicly accessible at the following URL: https://github.com/jamycheung/Trans4Trans.

Flying Guide Dog Concept [201]. In the pursuit of a mobility assistance system that offers spatially flexible interaction, we have devised a groundbreaking prototype known as the "flying guide dog". This innovative system combines drone technology with advanced scene understanding capabilities to provide a novel user experience. By executing semantic segmentation on the frames captured by the drone's camera, contextual visual information like sidewalks, crosswalks, and traffic lights is extracted from the environment. With this environmental perception in hand, the drone can adeptly adjust its own trajectory, guiding the user to navigate their surroundings safely. In practical terms, the user's interaction with the system involves holding a soft string that is affixed to the drone. This tethered connection ensures that the user can easily follow the drone's movements. Additionally, the user receives pertinent voice prompts through a Bluetooth-enabled bone conduction headphone, further enhancing their awareness of their environment and the drone's guidance cues. The system implementation have been made publicly accessible at the following URL: https://github.com/EckoTan0804/flying-guide-dog.

9 OUTLOOK

In this chapter, we post a discussion and look towards the future of intelligent transportation and mobility assistance systems. Starting from the ideas for improving assistance systems, we outline potential iterative developments for wearable systems by actively involving users in the refinement process. Furthermore, we introduce the promising prospects of harnessing large language models and vision-language models to shape the landscape of smart mobility.

9.1 NEXT-GENERATION ASSISTANCE SYSTEM

Based on the insightful feedback received from expert participants in our Vision4Blind system studies, several key discussions emerged regarding the optimal delivery of information to users. Notably, four out of the five experts have expressed the necessity for distance information related to recognized objects, while one expert underscored the significance of clock directions or stereo sound cues to aid object localization. They highlighted the importance of presenting information about objects directly in front of the user, as opposed to those positioned to the side, which would significantly elevate both user experience and accuracy. Furthermore, three experts have engaged in discussions around scenarios where grasping the spatial arrangement of objects is critical, particularly in distinguishing between items like tables and windows. Their input advocated for the communication of object positions in a sequence, potentially incorporating levels or layers to effectively convey the hierarchy of distances. In aggregate, the expert feedback reinforce the imperative of providing precise, contextually enriched information about object directions and distances, thereby enhancing the scene understanding. Leveraging this invaluable feedback, the next-generation assistance system can be improved to provide object distancing and orientation enhancements, as well as a more efficient user interface to convey information.

9.2 VISION-LANGUAGE MODELS FOR INTERACTIVE NAVIGATION

Vision-Language Navigation (VLN) represents a novel research direction that has potential to transform the landscape of navigation systems. At its core, VLN explores the challenge of enabling a mobile agent to traverse unfamiliar environments by comprehending and following textual instructions provided by an oracle. This task is a significant stride towards empowering machines with the ability to navigate and interact with the real world. The fundamental essence of VLN lies in its capacity to surmount the absence of prior knowledge about the environment. Unlike traditional navigation systems that rely on predefined maps, VLN forges ahead by relying solely on textual guidance to maneuver through unfamiliar terrain.

The insights and advancements garnered from VLN research can pave the way for the development of navigation systems across various domains. For instance, VLN principles can be seamlessly integrated into autonomous vehicles, imbuing them with the capability to navigate complex urban environments. Similarly, drones and robots can leverage VLN techniques to navigate through cluttered spaces or hazardous terrains. One of the most transformative applications of VLN is its potential to serve as a bedrock for interactive navigation systems tailored to assist people with visual impairments. By combining natural language understanding with visual perception, VLN can enable real-time communication between the user and the wearable assistive system. Such systems can provide detailed step-by-step instructions, contextual cues, and real-time feedback, empowering people with visual impairments to navigate their surroundings with greater confidence and independence.

9.3 A UNIVERSAL MODEL FOR MULTIPLE TASKS

A promising research direction of multiple learning lies in using a unified vision-language model to address an array of interrelated tasks. It offers several benefits, including the consolidation of multiple tasks under a single model, streamlining development efforts, and improving system efficiency. Traditionally, various vision-relevant tasks such as traffic light recognition, scene segmentation, object identification, and zebra-crossing recognition entail the deployment of disparate models. Using a unified vision-language model enables the seamless integration of these tasks, resulting in significant reduction of development complexity and computational overhead. Furthermore, the unified model extends its versatility to solve vision-language tasks. For instance, this model can perform vision-language navigation, tackle open-vocabulary segmentation, and also handle visual question answering, responding to user queries about their surroundings. This unified approach holds immense promise for the future of intelligent transportation and mobility assistance. By integrating multiple tasks under one model, development efforts are reduced, computational resources are optimized, and the system efficiency is enhanced.

9.4 LARGE MODELS IN ASSISTIVE TECHNOLOGY

Artificial General Intelligence (AGI) holds the potential to drive the advancement of assistive technology across various domains. This includes the fields of scene perception and understanding, scene reasoning, adaptive learning tailored to individual user preferences and needs, as well as high-level decision-making in the fields of mobility and navigation. These capabilities can be used for developing smarter and more adaptable assistive technologies. However, the current AGI technology comes with significant resource demands. As a result, ensuring the accessibility and affordability of AGI-powered assistance systems to a broad spectrum of users presents a substantial challenge. Therefore, a critical aspect lies in enhancing the efficiency of AGI-based systems, a factor that directly contributes to enhancing the overall user experience.

Part V

APPENDIX

A

APPENDIX

A.1 MORE INFORMATION OF DELIVER DATASET

A.1.1 *Data Collection*

Depth2Frames. The depth camera straightforwardly outputs a grayscale depth map (*i.e.* 0–255 scales), which will cause discontinuity and quantization errors in distance measurements. We convert the original depth image to the depth frame using a logarithmic scale, leading to milimetric granularity and better precision at close ranges.
Event2Frames. The positive- and negative event threshold of the event camera are both set to 0.3. We record raw event point cloud between two adjacent frames and convert the last occurring event among all pixels into an event frame, where blue indicates positive and red indicates negative.
LiDAR2Frames. We transform the LiDAR point cloud to the image coordinate system, so as to obtain an image-like representation of LiDAR data. The Field-of-View (FoV) of the front camera is 91° and the image resolution is $H\times W{=}1042\times 1042$. The origin is $(u_0, v_0){=}(H/2, W/2)$. The focal length (f_x, f_y) is calculated as:

$$f_x{=}H/(2\times\tan(FoV\times\pi/360)), \tag{40}$$

$$f_y{=}W/(2\times\tan(FoV\times\pi/360)). \tag{41}$$

To project 3D points to 2D image coordinate, we have:

$$\begin{bmatrix}u\\ v\\ 1\end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0\\ 0 & f_y & v_0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\\ \mathbf{o}_{3\times 1}^{T} & 1\end{bmatrix}\begin{bmatrix}X\\ Y\\ Z\\ 1\end{bmatrix}, \tag{42}$$

where (X, Y, Z) is the LiDAR point, (u, v) is the 2D image pixel, and the rotation ($\boldsymbol{R}$) and the translation ($\boldsymbol{t}$) matrices are set as the unit matrix in the CARLA simulator [53].

A.1.2 *Dataset Structure*

DeLiVER contains Depth, LiDAR, Event, and RGB modalities. As shown in Figure 68, four adverse road scene conditions of *rainy*, *sunny*, *foggy*, and *night* are included in our dataset. There are five sensor failure cases including Motion Blur (**MB**), Over-Exposure (**OE**), Under-Exposure (**UE**), LiDAR-Jitter (**LJ**), and Event Low-resolution (**EL**) to verify that the performance of model is robust and stable in the presence of sensor failures. The sensors are mounted at different locations on the ego car to provide multiple views

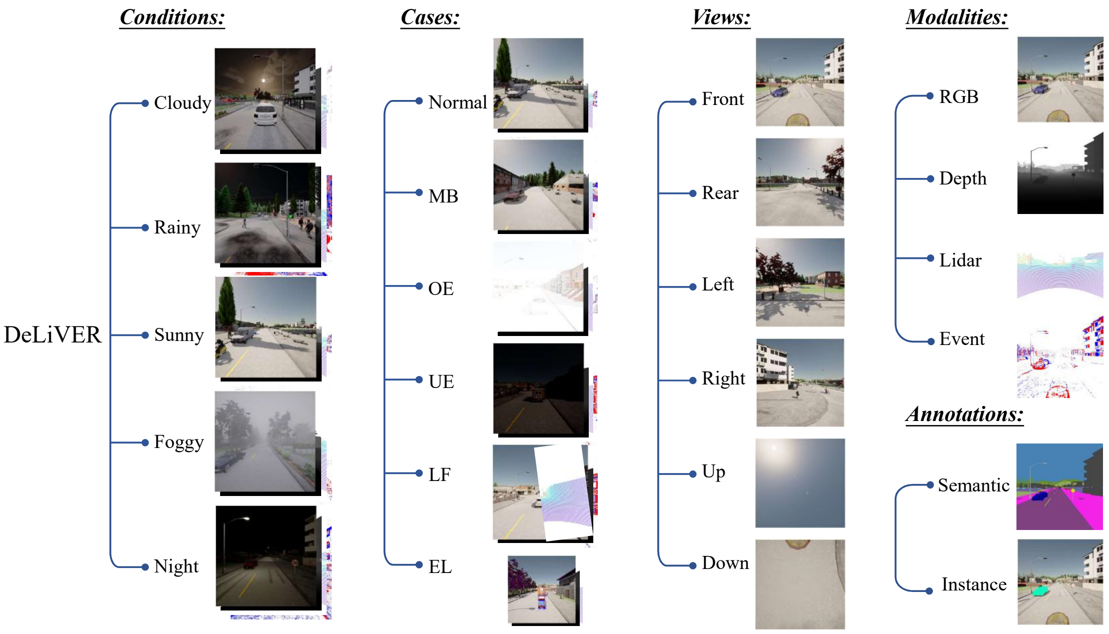


Figure 68: Data structure of the DeLiVER dataset. The columns from left to right are respective conditions, cases, multiple views, modalities and annotations. **MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution.

Split	**Cloudy**	**Foggy**	**Night**	**Rainny**	**Sunny**	**Total**	**Normal**	**MB**	**OE**	**UE**	**LJ**	**EL**	**Total**
Train	794	795	797	799	798	3983	2585	600	200	199	199	200	3983
Val	398	400	410	398	399	2005	1298	299	100	99	100	109	2005
Test	379	379	379	380	380	1897	1198	300	100	100	99	100	1897
Front-view	1571	1574	1586	1577	1577	7885	5081	1199	400	398	398	409	7885
All six views	9426	9444	9516	9462	9462	47310	30486	7194	2400	2388	2388	2454	47310

Table 34: Data statistic of DeLiVER dataset. It includes four adverse conditions (*cloudy*, *foggy*, *rainy*, and *night*), and each condition has five failure cases (**MB**: Motion Blur; **OE**: Over-Exposure; **UE**: Under-Exposure; **LJ**: LiDAR-Jitter; and **EL**: Event Low-resolution).

including *front*, *rear*, *left*, *right*, *up*, and *down*. Each sample is annotated with semantic and instance labels. In this work, we focus on the front-view semantic segmentation.

There are 25 classes in DeLiVER dataset: *Building, Fence, Other, Pedestrian, Pole, RoadLine, Road, SideWalk, Vegetation, Cars, Wall, TrafficSign, Sky, Ground, Bridge, RailTrack, GroundRail, TrafficLight, Static, Dynamic, Water, Terrain, TwoWheeler, Bus, Truck.*

A.1.3 *Dataset Statistic*

We present statistics of the DeLiVER dataset in Table 34. We discuss data partitioning in two groups, one according to the conditions and the other according to the sensor failures. Note that, the two groups are mutually inclusive. The five cases from the second group are included in each of five conditions from the first group. For example, cases of **MB**, **OE**, **UE**, **LJ**, and **EL** are included in *cloudy*, *foggy*, *night*, *rainy*, and *sunny* conditions, but with different samples. To investigate the robustness under sensor failures, we collect 1199, 400, 398, 398, and 409 frames on respective cases.

A.2 DRONE CONTROL ALGORITHM

The control strategy elaborated in Section 7.2 is detailed in Algorithm 2. This algorithm shows the intricate steps and decision-making processes of the drone, acting as a flying guide dog for individuals with visual impairments.

Algorithm 2 : Drone control

```
input  : colorized prediction $C \in \mathbb{R}^{H \times W \times 3}$, pedestrian crossing light
         color $\in \{red, green, None\}$,
         current altitude $h$
output : up/down velocity $v_{ud}$, yaw velocity $v_{yaw}$, left/right velocity $v_{lr}$,
         forward velocity $v_f$

Initialize parameters: pixel threshold $\theta_{conf}$, target altitude $h_{target}$, speed up
  speedup, up/down velocity $v_{ud}$, forward velocity $v_{f,0}$; lists: preset yaw
  velocities $list_{yaws}$, binary control code $list_{codes}$;
start_crossing ← false;
while drone is flying and video stream is on do
    // Maintain target altitude
    if $h \neq h_{target}$ then
        $v_{ud} \leftarrow (h_{target} - h)/h_{target} * v_{ud}$;
    end
    // Fly along walkable path
    Extract largest walkable area $L_{walkable}$ from $C$;
    $(x_{centroid}, y_{centroid}) \leftarrow$ estimate_centroid$(L_{walkable})$ ;
    $v_{lr} \leftarrow x_{centroid} - \frac{W}{2}$;
    $(R_l, R_m, R_r) \leftarrow$ partition$(L_{walkable})$;
    for $p \in (R_l, R_m, R_r)$ do
        conf ← mean(p);
        code ← binary_conversion(conf, $\theta_{conf}$);
        Append code to $list_{codes}$;
    end
    $v_{yaw} \leftarrow$ get_yaw_vel$(list_{yaws}, list_{codes})$;
    // Fusion with traffic light classification
    if color = None then
        $v_f \leftarrow v_{f,0}$;
    else if color = green or start_crossing then
        $v_f \leftarrow v_{f,0} +$ speedup;
        start_crossing ← true;
    else
        $v_f \leftarrow 0$;
    end
end
```

B

PUBLICATIONS

This doctoral research resulted in the following peer-reviewed publications, which are incorporated in whole or in part in this thesis. (* equal contribution, † corresponding.)

[1] J. Zhang*, R. Liu*, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, R. Stiefelhagen. **Delivering Arbitrary-Modal Semantic Segmentation.** In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. Stiefelhagen. **Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation.** In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] J. Zhang*, H. Liu*, K. Yang*, X. Hu, R. Liu, and R. Stiefelhagen. **CMX: Cross-modal Fusion for RGB-X Semantic Segmentation with Transformers.** IEEE *Transactions on Intelligent Transportation Systems (T-ITS)*, 2023.

[4] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen. **Trans4Trans: Efficient Transformer for Transparent Object and Semantic Scene Segmentation in real-world Navigation Assistance.** IEEE *Transactions on Intelligent Transportation Systems (T-ITS)*, 2022.

[5] J. Zhang, C. Ma, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen. **Transfer beyond the Field of View: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation.** IEEE *Transactions on Intelligent Transportation Systems (T-ITS)*, 2021.

[6] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen. **Trans4Trans: Efficient Transformer for Transparent Object Segmentation to help Visually Impaired People Navigate in the real world.** In *International Workshop on Assistive Computer Vision and Robotics (ACVR) with IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[7] Z. Teng*, J. Zhang*†, K. Yang, K. Peng, H. Shi, S. Reiß, K. Cao, R. Stiefelhagen. **360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View.** In *Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[8] C. Chen, J. Zhang†, K. Yang, K. Peng, and R. Stiefelhagen. **Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping from Egocentric Images to Allocentric Semantics with Vision Transformers.** In *Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[9] Q. Wang*, J. Zhang*, K. Yang, K. Peng, and R. Stiefelhagen. **MatchFormer: Interleaving Attention in Transformers for Feature Matching.** In *Asian Conference on Computer Vision (ACCV)*, 2022.

[10] X. Luo*, J. Zhang*, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen. **Towards Robust Semantic Segmentation of Accident Scenes via Multi-Source Mixed Sampling and Meta-Learning.** In *Workshop on Autonomous Driving (WAD) with IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] H. Tan, C. Chen, X. Luo, J. Zhang, C. Seibold, K. Yang, R. Stiefelhagen. **Flying Guide Dog: Walkable Path Discovery for the Visually Impaired Utilizing Drones and Transformer-based Semantic Segmentation.** IEEE *International Conference on Robotics and Biomimetics (ROBIO)*, 2021.

The following publications were co-authored by Jiaming Zhang, but not included in this thesis. (in chronological order, * equal contribution, † corresponding.)

[12] O. Moured, J. Zhang, A. Roitberg, T. Schwarz, and R. Stiefelhagen. **Line Graphics Digitization: A Step Towards Full Automation.** In *International Conference on Document Analysis and Recognition (ICDAR)*, 2023.

[13] R. Liu, J. Zhang†, K. Peng, J. Zheng, K. Cao, Y. Chen, K. Yang, R. Stiefelhagen. **Open Scene Understanding: Grounded Situation Recognition Meets Segment Anything for Helping People with Visual Impairments.** In *International Workshop on Assistive Computer Vision and Robotics (ACVR) with IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] F. Teng*, J. Zhang*, K. Peng, K. Yang, Y. Wang, and R. Stiefelhagen. **OAFuser: Towards Omni-Aperture Fusion for Light Field Semantic Segmentation of Road Scenes.** In arXiv preprint arXiv:2307.15588, 2023. **Under review.**

[15] K. Cao, R. Liu, Z. Wang, K. Peng, J. Zhang, J. Zheng, Z. Teng, K. Yang, R. Stiefelhagen. **Tightly-Coupled LiDAR-Visual SLAM Based on Geometric Features for Mobile Agents.** IEEE *International Conference on Robotics and Biomimetics (ROBIO)*, 2023.

[16] K. Peng, D. Wen, D. Schneider, J. Zhang, K. Yang, M.S. Sarfraz, R. Stiefelhagen, A. Roitberg. **FeatFSDA: Towards Few-shot Domain Adaptation for Video-based Activity Recognition.** In arXiv preprint arXiv:2305.08420, 2023. **Under review.**

[17] S. Li, K. Yang, H. Shi, J. Zhang, J. Lin, Z. Teng, and Z. Li. **Bi-Mapper: Holistic BEV Semantic Mapping for Autonomous Driving.** IEEE *Robotics and Automation Letters*, 2023.

[18] H. Shi, Y. Li, K. Yang, J. Zhang, K. Peng, A. Roitberg, Y. Ye, H. Ni, K. Wang, R. Stiefelhagen. **FishDreamer: Towards Fisheye Semantic Completion via Unified Image Outpainting and Segmentation.** In *Omnidirectional Computer Vision (OmniCV) Workshop with IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] J. Zheng*, J. Zhang*, K. Yang, K. Peng, and R. Stiefelhagen. **MATERobot: Material Recognition in Wearable Robotics for People with Visual Impairments.** In arXiv preprint arXiv:2302.14595, 2023. **Under review.**

[20] K. Peng, D. Schneider, A. Roitberg, K. Yang, J. Zhang, M.S. Sarfraz, R. Stiefelhagen. **MuscleMap: Towards Video-based Activated Muscle Group Estimation.** In arXiv preprint arXiv:2303.00952, 2023. **Under review.**

[21] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. **Multi-modal depression estimation based on sub-attentional fusion.** In *International Workshop on Assistive Computer Vision and Robotics (ACVR) with European Conference on Computer Vision (ECCV)*, 2022.

[22] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. **Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions.** IEEE *Transactions on Multimedia (TMM)*, 2023.

[23] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. **TransDARC: Transformer-based Driver Activity Recognition with Latent Space Feature Calibration.** In IEEE/RSJ *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[24] J. Zhang, K. Yang, H. Shi, S. Reiß, K. Peng, C. Ma, H. Fu, P. Torr, K. Wang, R. Stiefelhagen. **Behind Every Domain There is a Shift: Adapting Distortion-aware Vision Transformers for Panoramic Semantic Segmentation.** In arXiv preprint arXiv:2207.11860, 2022. **Under review.**

[25] W. Ou, J. Zhang, K. Peng, K. Yang, G. Jaworek, K. Müller, R. Stiefelhagen. **Indoor navigation assistance for visually impaired people via dynamic SLAM and panoptic segmentation with an RGB-D sensor.** In *International Conference on Computers Helping People with Special Needs (ICCHP)*, 2022.

[26] R. Liu, K. Yang, A. Roitberg, J. Zhang, K. Peng, H. Liu, R. Stiefelhagen. **TransKD: Transformer knowledge distillation for efficient semantic segmentation.** In arXiv preprint arXiv:2202.13393, 2022. **Under review.**

[27] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. **ProFormer: Learning Data-efficient Representations of Body Movement with Prototype-based Feature Augmentation and Visual Transformers.** In arXiv preprint arXiv:2202.11423, 2022. **Under review.**

[28] K. Peng, J. Fei, K. Yang, A. Roitberg, J. Zhang, F. Bieder, P. Heidenreich, C. Stiller, R. Stiefelhagen. **MASS: Multi-attentional semantic segmentation of LiDAR data for dense top-view understanding.** IEEE *Transactions on Intelligent Transportation Systems (T-ITS)*, 2022.

[29] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen. **Should I take a walk? Estimating Energy Expenditure from Video Data.** In *International Workshop on Computer Vision for Physiological Measurement (CVPM) with IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] R. Liu, K. Yang, H. Liu, J. Zhang, K. Peng, and R. Stiefelhagen. **Transformer-based knowledge distillation for efficient semantic segmentation of road-driving scenes.** In arXiv preprint arXiv:2202.13393, 2022. **Under review.**

[31] J. Zhang, K. Yang, and R. Stiefelhagen. **Exploring Event-driven Dynamic Context for Accident Scene Segmentation.** IEEE *Transactions on Intelligent Transportation Systems (T-ITS)*, 2021.

[32] J. Zhang, K. Yang, and R. Stiefelhagen. **ISSAFE: Improving Semantic Segmentation in Accidents by Fusing Event-based Data.** In IEEE/RSJ *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[33] C. Ma, J. Zhang, K. Yang, A. Roitberg, and R. Stiefelhagen. **DensePASS: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation with Attention-Augmented Context Exchange.** IEEE *International Intelligent Transportation Systems Conference (ITSC)*, 2021.

[34] Z. Marinov, S. Vasileva, Q. Wang, C. Seibold, J. Zhang, and R. Stiefelhagen. **Pose2Drone: A Skeleton-Pose-based Framework forHuman-Drone Interaction.** In *European Signal Processing Conference (EUSIPCO)*, 2021.

[35] W. Mao, J. Zhang, K. Yang, and R. Stiefelhagen. **Panoptic Lintention Network: Towards Efficient Navigational Perception for the Visually Impaired.** IEEE *International Conference on Real-Time Computing and Robotics (RCAR)*, 2021.

[36] Y. Zhang, H. Chen, K. Yang, J. Zhang, and R. Stiefelhagen. **Perception Framework through Real-Time Semantic Segmentation and Scene Recognition on a Wearable System for the Visually Impaired.** IEEE *International Conference on Real-Time Computing and Robotics (RCAR)*, 2021.

[37] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen. **Capturing Omni-Range Context for Omnidirectional Segmentation.** In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[38] H. Liu, R. Liu, K. Yang, J. Zhang, K. Peng, and R. Stiefelhagen. **HIDA: Towards Holistic Indoor Understanding for the Visually Impaired via Semantic Instance Segmentation with a Wearable Solid-State LiDAR Sensor.** In *International Workshop on Assistive Computer Vision and Robotics (ACVR) with IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[39] W. Mao, J. Zhang, K. Yang, and R. Stiefelhagen. **Can we cover navigational perception needs of the visually impaired by panoptic segmentation?.** In arXiv preprint arXiv:2007.10202, 2020.

BIBLIOGRAPHY

[1] Shivendra Agrawal, Mary Etta West, and Bradley Hayes. "A Novel Perceptive Robotic Cane with Haptic Navigation for Enabling Vision-Independent Participation in the Social Dynamics of Seat Choice." In: *IROS*. 2022.

[2] Aitor Aladren, Gonzalo López-Nicolás, Luis Puig, and Josechu J. Guerrero. "Navigation assistance for the visually impaired using RGB-D sensor with range expansion." In: *IEEE Systems Journal* (2016).

[3] Inigo Alonso and Ana C. Murillo. "EV-SegNet: Semantic segmentation for event-based cameras." In: *CVPRW*. 2019.

[4] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. "Joint 2D-3D-semantic data for indoor scene understanding." In: *arXiv preprint arXiv:1702.01105* (2017).

[5] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. "Joint 2D-3D-semantic data for indoor scene understanding." In: *arXiv preprint arXiv:1702.01105* (2017).

[6] Mauro Avila Soto, Markus Funk, Matthias Hoppe, Robin Boldt, Katrin Wolf, and Niels Henze. "DroneNavigator: Using leashed and free-floating quadcopters to navigate visually impaired travelers." In: *ASSETS*. 2017.

[7] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. "MultiMAE: Multi-modal Multi-task Masked Autoencoders." In: *ECCV*. 2022.

[8] Jinqiang Bai, Shiguo Lian, Zhaoxiang Liu, Kai Wang, and Dijun Liu. "Smart guiding glasses for visually impaired people in indoor environment." In: *IEEE Trans. Consumer Electron.* (2017).

[9] Federica Barontini, Manuel G Catalano, Lucia Pallottino, Barbara Leporini, and Matteo Bianchi. "Integrating wearable haptics and obstacle avoidance for the visually impaired in indoor navigation: A user-centered approach." In: *IEEE Transactions on Haptics* 14.1 (2020), pp. 109–122.

[10] Bruno Berenguel-Baeta, Manuel Guerrero-Viu, A. Nova, Jesus Bermudez-Cameo, Alejandro Pérez-Yus, and Josechu J. Guerrero. "Floor Extraction and Door Detection for Visually Impaired Guidance." In: *ICARCV*. 2020.

[11] Julie Stephany Berrio, Mao Shan, Stewart Worrall, and Eduardo Nebot. "Camera-LIDAR integration: Probabilistic sensor fusion for semantic mapping." In: *T-ITS* (2021).

[12] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence." In: *CVPR*. 2017.

[13] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology." In: *Qualitative Research in Psychology* (2006).

[14] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. "HRFuser: A Multi-resolution Sensor Fusion Architecture for 2D Object Detection." In: *arXiv preprint arXiv:2206.15157* (2022).

[15] Matthew J Burton, Jacqueline Ramke, Ana Patricia Marques, Rupert RA Bourne, Nathan Congdon, Iain Jones, Brandon AM Ah Tong, Simon Arunga, Damodar Bachani, Covadonga Bascaran, et al. "The Lancet global health Commission on global eye health: vision beyond 2020." In: *The Lancet Global Health* 9.4 (2021), e489–e551.

[16] F. M. Butera. "Glass architecture: is it sustainable." In: *Passive and Low Energy Cooling for the Built Environment* (2005).

[17] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. "COCO-Stuff: Thing and Stuff Classes in Context." In: *CVPR*. 2018.

[18] Yuqi Cai, Wujie Zhou, Liting Zhang, Lu Yu, and Ting Luo. "DHFNet: Dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation." In: *The Visual Computer* 1 (2023), pp. 1–11.

[19] Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, and Luc Van Gool. "Structured Bird's-Eye-View Traffic Scene Understanding From Onboard Images." In: *CVPR*. 2021.

[20] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. "ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation." In: *ICCV*. 2021.

[21] Zhengcai Cao, Xiaowen Xu, Biao Hu, and MengChu Zhou. "Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network." In: *IEEE Trans. Intell. Transp. Syst.* (2021).

[22] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. "Semantic MapNet: Building Allocentric Semantic Maps and Representations from Egocentric Views." In: *AAAI*. 2021.

[23] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. "Matterport3D: Learning from RGB-D Data in Indoor Environments." In: *3DV*. 2017.

[24] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. "HarDNet: A low memory traffic network." In: *ICCV*. 2019.

[25] Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. "Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping From Egocentric Images to Allocentric Semantics With Vision Transformers." In: *WACV*. 2023.

[26] Haoye Chen, Yingzhi Zhang, Kailun Yang, Manuel Martinez, Karin Müller, and Rainer Stiefelhagen. "Can We Unify Perception and Localization in Assisted Navigation? An Indoor Semantic Visual Positioning System for Visually Impaired People." In: *ICCHP*. 2020.

[27] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. "Dynamic region-aware convolution." In: *CVPR*. 2021.

[28] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2018).

[29] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In: *ECCV*. 2018.

[30] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In: *ECCV*. 2018.

[31] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. "Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation." In: *TIP* 30 (2021), pp. 2313–2324.

[32] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. "Spatial information guided convolution for real-time RGBD semantic segmentation." In: *TIP* (2021).

[33] Long Chen et al. "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning." In: *CVPR*. 2017.

[34] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. "CycleMLP: A MLP-like architecture for dense prediction." In: *ICLR*. 2022.

[35] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. "Big Self-Supervised Models are Strong Semi-Supervised Learners." In: *NeurIPS*. 2020.

[36] Xiaokang Chen et al. “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation.” In: *ECCV.* 2020.

[37] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. “DPT: Deformable Patch-based Transformer for Visual Recognition.” In: *MM.* 2021.

[38] Zhuo Chen, Xiaoming Liu, Masaru Kojima, Qiang Huang, and Tatsuo Arai. “A wearable navigation device for visually impaired people based on the real-time semantic visual SLAM system.” In: *Sensors* (2021).

[39] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. “Per-Pixel Classification is Not All You Need for Semantic Segmentation.” In: *NeurIPS.* 2021.

[40] Ruiqi Cheng, Kaiwei Wang, Jian Bai, and Zhijie Xu. “Unifying visual localization and scene recognition for people with visual impairment.” In: *IEEE Access* (2020).

[41] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. “Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation.” In: *CVPR.* 2017.

[42] Sungha Choi, Joanne T. Kim, and Jaegul Choo. “Cars Can’t Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks.” In: *CVPR.* 2020.

[43] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. “Gauge equivariant convolutional networks and the icosahedral CNN.” In: *ICML.* 2019.

[44] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding.” In: *CVPR.* 2016.

[45] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. “ScanNet: Richly-annotated 3D reconstructions of indoor scenes.” In: *CVPR.* 2017.

[46] Angela Dai and Matthias Nießner. “3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation.” In: *ECCV.* 2018.

[47] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. “Deformable convolutional networks.” In: *ICCV.* 2017.

[48] Fuqin Deng et al. “FEANet: Feature-enhanced Attention Network for RGB-thermal Real-time Semantic Segmentation.” In: *IROS.* 2021.

[49] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. “Restricted deformable convolution-based road scene semantic segmentation using surround view cameras.” In: *T-ITS* 21.10 (2020), pp. 4350–4362.

[50] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. “SuperPoint: Self-supervised interest point detection and description.” In: *CVPRW.* 2018.

[51] Xiaoyi Dong et al. “CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows.” In: 2021.

[52] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” In: *ICLR.* 2021.

[53] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. “CARLA: An open urban driving simulator.” In: *CoRL.* 2017.

[54] Ping-Jung Duh, Yu-Cheng Sung, Liang-Yu Fan Chiang, Yung-Ju Chang, and Kuan-Wen Chen. “V-Eye: A Vision-based Navigation System for the Visually Impaired.” In: *IEEE Trans. Multim.* (2021).

[55] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. “D2-net: A trainable CNN for joint detection and description of local features.” In: *CVPR* (2019).

[56] Pramit Dutta, Ganesh Sistu, Senthil Yogamani, Edgar Galván, and John McDonald. “ViT-BEVSeg: A Hierarchical Transformer Network for Monocular Birds-Eye-View Segmentation.” In: *IJCNN.* 2022.

[57] Isht Dwivedi, Srikanth Malla, Yi-Ting Chen, and Behzad Dariush. “Bird’s eye view segmentation using lifted 2D semantic features.” In: *BMVC*. 2021.

[58] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. “Tangent images for mitigating spherical distortion.” In: *CVPR*. 2020.

[59] Fahimeh Fooladgar and Shohreh Kasaei. “Multi-modal attention-based fusion model for semantic segmentation of RGB-depth images.” In: *arXiv preprint arXiv:1912.11691* (2019).

[60] Oriel Frigo, Lucien Martin-Gaffé, and Catherine Wacongne. “DooDLeNet: Double DeepLab Enhanced Feature Fusion for Thermal-color Semantic Segmentation.” In: *CVPRW*. 2022.

[61] Jun Fu et al. “Dual attention network for scene segmentation.” In: *CVPR*. 2019.

[62] Guillermo Gallego et al. “Event-based vision: A survey.” In: *TPAMI* (2022).

[63] Biao Gao, Yancheng Pan, Chengkun Li, Sibo Geng, and Huijing Zhao. “Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods.” In: *T-ITS* (2022).

[64] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. “Combining Events and Frames Using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction.” In: *RA-L* 6.2 (2021), pp. 2822–2829.

[65] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? The KITTI vision benchmark suite.” In: *CVPR*. 2012.

[66] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. “Omnivore: A Single Model for Many Visual Modalities.” In: *CVPR*. 2022.

[67] Nikhil Gosala and Abhinav Valada. “Bird’s-eye-view panoptic segmentation using monocular frontal view images.” In: *RA-L* (2022).

[68] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. “Volumetric Instance-Aware Semantic Mapping and 3D Object Discovery.” In: *RA-L* (2019).

[69] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. “Multi-scale high-resolution vision transformer for semantic segmentation.” In: *CVPR*. 2022.

[70] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. “SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation.” In: *NeurIPS*. 2022.

[71] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. “MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes.” In: *IROS*. 2017.

[72] Sandra G. Hart. “NASA-Task Load Index (NASA-TLX); 20 Years Later.” In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2006).

[73] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. “FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture.” In: *ACCV*. 2016.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In: *CVPR*. 2016.

[75] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. “Vision permutator: A permutable MLP-like architecture for visual recognition.” In: *TPAMI* 45.1 (2023), pp. 1328–1334.

[76] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. “Strip pooling: Rethinking spatial pooling for scene parsing.” In: *CVPR*. 2020.

[77] I-Hsuan Hsieh, Hsiao-Chu Cheng, Hao-Hsiang Ke, Hsiang-Chieh Chen, and Wen-June Wang. “Outdoor walking guide for the visually-impaired people based on semantic segmentation and depth map.” In: *ICPAI*. 2020.

[78] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks.” In: *CVPR*. 2018.

[79] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe Lin, Kate Saenko, and Stan Sclaroff. “Real-time semantic segmentation with fast attention.” In: *RA-L* 6.1 (2021), pp. 263–270.

[80] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. “Bidirectional projection network for cross dimension scene understanding.” In: *CVPR*. 2021.

[81] Xing Hu, Yi An, Cheng Shao, and Huosheng Hu. “Distortion Convolution Module for Semantic Segmentation of Panoramic Images Based on the Image-Forming Principle.” In: *TIM* 71 (2022), pp. 1–12.

[82] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. “ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation.” In: *ICIP*. 2019.

[83] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. “The ApolloScape open dataset for autonomous driving and its application.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[84] Zhiming Huang, Kaiwei Wang, Kailun Yang, Ruiqi Cheng, and Jian Bai. “Glass detection and recognition based on the fusion of ultrasonic sensor and RGB-D sensor for the visually impaired.” In: *SPIE*. 2018.

[85] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. “CCNet: Criss-cross attention for semantic segmentation.” In: *ICCV*. 2019.

[86] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Nießner. “Spherical CNNs on Unstructured Grids.” In: *ICLR*. 2019.

[87] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. “RedNet: Residual encoder-decoder network for indoor rgb-d semantic segmentation.” In: *arXiv preprint arXiv:1806.01054* (2018).

[88] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. “COTR: Correspondence Transformer for Matching Across Images.” In: *ICCV*. 2021.

[89] Qiangguo Jin, Zhaopeng Meng, Tuan D. Pham, Qi Chen, Leyi Wei, and Ran Su. “DUNet: A deformable network for retinal vessel segmentation.” In: *Knowl. Based Syst.* (2019).

[90] Zhenchao Jin et al. “Mining Contextual Information Beyond Image for Semantic Segmentation.” In: *ICCV*. 2021.

[91] Christopher J. Johnstone, Nicole A. Bottsford-Miller, and Sandra J. Thompson. *Using the Think Aloud Method (Cognitive Labs) To Evaluate Test Design for Students with Disabilities and English Language Learners*. Tech. rep. National Center on Educational Outcomes, University of Minnesota, 2006.

[92] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. “MMTM: Multimodal transfer module for CNN fusion.” In: *CVPR*. 2020.

[93] Esteban Bayro Kaiser and Michael Lawo. “Wearable navigation system for the visually impaired and blind people.” In: *2012 IEEE/ACIS 11th International Conference on Computer and Information Science*. IEEE. 2012, pp. 230–233.

[94] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C. V. Jawahar. “Universal Semi-Supervised Semantic Segmentation.” In: *ICCV*. 2019.

[95] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. “Deep polarization cues for transparent object segmentation.” In: *CVPR*. 2020.

[96] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. “Intel RealSense stereoscopic depth cameras.” In: *CVPRW*. 2017.

[97] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. “Panoptic feature pyramid networks.” In: *CVPR*. 2019.

[98] Shu Kong and Charless C. Fowlkes. “Recurrent scene parsing with perspective understanding in the loop.” In: *CVPR*. 2018.

[99] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. "SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360°Images." In: *CVPR*. 2019.

[100] Gen Li, Inyoung Yun, Jonghyun Kim, and Joongkyu Kim. "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation." In: *BMVC*. 2019.

[101] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. "RGB-T Semantic Segmentation with Location, Activation, and Sharpening." In: *TCSVT* (2022).

[102] Guoxin Li, Jiaqi Xu, Zhijun Li, Chao Chen, and Zhen Kan. "Sensing and navigation of wearable assistance cognitive systems for the visually impaired." In: *TCDS* (2023).

[103] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. "DFANet: Deep feature aggregation for real-time semantic segmentation." In: *CVPR*. 2019.

[104] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. "ConvMLP: Hierarchical Convolutional MLPs for Vision." In: *arXiv preprint arXiv:2109.04454* (2021).

[105] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. "Situation recognition with graph neural networks." In: *ICCV*. 2017.

[106] Xiang Li, Hanzhang Cui, John-Ross Rizzo, Edward Wong, and Yi Fang. "Cross-Safe: A computer vision-based approach to make all intersection-related pedestrian signals accessible for the visually impaired." In: *CVC*. 2020.

[107] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. "Dual-resolution correspondence networks." In: *NeurIPS* (2020).

[108] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. "Contextual transformer networks for visual recognition." In: *TPAMI* 45.2 (2023), pp. 1489–1500.

[109] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. "Bidirectional learning for domain adaptation of semantic segmentation." In: *CVPR*. 2019.

[110] Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. "CTNet: Context-based tandem network for semantic segmentation." In: *TPAMI* (2021).

[111] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers." In: *ECCV*. 2022.

[112] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. "AS-MLP: An axial shifted MLP architecture for vision." In: *ICLR*. 2022.

[113] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. "Multimodal Material Segmentation." In: *CVPR*. 2022.

[114] Yiyi Liao, Jun Xie, and Andreas Geiger. "KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D." In: *TPAMI* 45.3 (2023), pp. 3292–3310.

[115] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. "Cascaded Feature Network for Semantic Segmentation of RGB-D Images." In: *ICCV*. 2017.

[116] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation." In: *CVPR*. 2017.

[117] Jiaying Lin, Zebang He, and Rynson W. H. Lau. "Rich Context Aggregation With Reflection Prior for Glass Surface Detection." In: *CVPR*. 2021.

[118] Shufei Lin, Kaiwei Wang, Kailun Yang, and Ruiqi Cheng. "KrNet: A kinetic real-time convolutional neural network for navigational assistance." In: *ICCHP*. 2018.

[119] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In: *CVPR*. 2017.

[120] Yimin Lin, Kai Wang, Wanxin Yi, and Shiguo Lian. "Deep learning based wearable assistive system for visually impaired people." In: *ICCVW*. 2019.

[121] Huayao Liu, Ruiping Liu, Kailun Yang, Jiaming Zhang, Kunyu Peng, and Rainer Stiefelhagen. "HIDA: Towards Holistic Indoor Understanding for the Visually Impaired via Semantic Instance Segmentation with a Wearable Solid-State LiDAR Sensor." In: *ICCVW*. 2021.

[122] Mengyi Liu, Shuhui Wang, Yulan Guo, Yuan He, and Hui Xue. "Pano-SfMLearner: Self-Supervised Multi-Task Learning of Depth and Semantics in Panoramic Videos." In: *SPL* 28 (2021), pp. 832–836.

[123] Mengyu Liu and Hujun Yin. "Feature pyramid encoding network for real-time semantic segmentation." In: *BMVC*. 2019.

[124] Yazhou Liu, Yuliang Chen, Pongsak Lasang, and Quansen Sun. "Covariance attention for semantic segmentation." In: *TPAMI* 44.4 (2022), pp. 1805–1818.

[125] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows." In: *ICCV*. 2021.

[126] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s." In: *CVPR*. 2022.

[127] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In: *CVPR*. 2015.

[128] David G. Lowe. "Distinctive image features from scale-invariant keypoints." In: *International Journal of Computer Vision* (2004).

[129] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks." In: *RA-L* (2019).

[130] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. "Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation." In: *CVPR*. 2019.

[131] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. "ContextDesc: Local descriptor augmentation with cross-modality context." In: *CVPR*. 2019.

[132] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. "ASLFeat: Learning local features of accurate shape and localization." In: *CVPR*. 2020.

[133] Xinyu Luo*, Jiaming Zhang*, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. "Towards robust semantic segmentation of accident scenes via multi-source mixed sampling and meta-learning." In: *CVPRW*. 2022.

[134] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. "DensePASS: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation with Attention-Augmented Context Exchange." In: *ITSC*. 2021.

[135] Yinan Ma, Qi Xu, Yue Wang, Jing Wu, Chengnian Long, and Yi-Bing Lin. "EOS: An efficient obstacle segmentation for blind guiding." In: *FGCS* (2023).

[136] Adriano Mancini, Emanuele Frontoni, and Primo Zingaretti. "Mechatronic system to help visually impaired users during walking and running." In: *IEEE Trans. Intell. Transp. Syst.* (2018).

[137] Roberto Manduchi and Sri Kurniawan. "Mobility-related accidents experienced by people with visual impairment." In: *AER Journal: Research and Practice in Visual Impairment and Blindness* (2011).

[138] Magdalena Maringer, Nico Hauck, and Ardeshir Mahdavi. "Suitability Evaluation of Visual Indicators on Glass Walls and Doors for Visually Impaired People." In: *Appl. Mech. Mater.* (2019).

[139] Manuel Martinez, Kailun Yang, Angela Constantinescu, and Rainer Stiefelhagen. "Helping the Blind to Get through COVID-19: Social Distancing Assistant Using Real-Time Semantic Segmentation on RGB-D Video." In: *Sensors* (2020).

[140] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. “ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network.” In: *CVPR*. 2019.

[141] Haiyang Mei et al. “Don’t hit me! Glass detection in real-world scenes.” In: *CVPR*. 2020.

[142] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes.” In: *ICCV*. 2017.

[143] Jakob Nielsen. “Estimating the number of subjects needed for a thinking aloud test.” In: *International Journal of Human-Computer Studies* (1994).

[144] Atsuro Okazawa, Tomoyuki Takahata, and Tatsuya Harada. “Simultaneous transparent and non-transparent object segmentation with multispectral scenes.” In: *IROS*. 2019.

[145] Semih Orhan and Yalin Bastanlar. “Semantic segmentation of outdoor panoramic images.” In: *SIVP* 16.3 (2022), pp. 643–650.

[146] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. “In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images.” In: *CVPR*. 2019.

[147] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. “In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images.” In: *CVPR*. 2019.

[148] Wenyan Ou, Jiaming Zhang, Kunyu Peng, Kailun Yang, Gerhard Jaworek, Karin Müller, and Rainer Stiefelhagen. “Indoor Navigation Assistance for Visually Impaired People via Dynamic SLAM and Panoptic Segmentation with an RGB-D Sensor.” In: *ICCHP-AAATE*. 2022.

[149] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. “Cross-view semantic segmentation for sensing surroundings.” In: *RA-L* (2020).

[150] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. “RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation.” In: *ICCV*. 2017.

[151] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. “ENet: A deep neural network architecture for real-time semantic segmentation.” In: *arXiv preprint arXiv:1606.02147* (2016).

[152] Andra Petrovai and Sergiu Nedevschi. “Semantic Cameras for 360-Degree Environment Perception in Automated Urban Driving.” In: *T-ITS* 23.10 (2022), pp. 17271–17283.

[153] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. “Seamless Scene Segmentation.” In: *CVPR*. 2019.

[154] Rudra P. K. Poudel, Ujwal Bonde, Stephan Liwicki, and Christopher Zach. “ContextNet: Exploring context and detail for semantic segmentation in real-time.” In: *BMVC*. 2018.

[155] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. “Fast-SCNN: Fast semantic segmentation network.” In: *BMVC*. 2019.

[156] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. “Multi-Modal Fusion Transformer for End-to-End Autonomous Driving.” In: *CVPR*. 2021.

[157] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. “3D graph neural networks for RGBD semantic segmentation.” In: *ICCV*. 2017.

[158] Yeqiang Qian, Liuyuan Deng, Tianyi Li, Chunxiang Wang, and Ming Yang. “Gated-Residual Block for Semantic Segmentation Using RGB-D Data.” In: *T-ITS* 23.8 (2022), pp. 11836–11844.

[159] Digvijay S Raghuvanshi, Isha Dutta, and RJ Vaidya. “Design and analysis of a novel sonar-based obstacle-avoidance system for the visually impaired and unmanned systems.” In: *ICES*. IEEE. 2014.

[160] Teng Ran, Liang Yuan, Jianbo Zhang, Dingxin Tang, and Li He. “RS-SLAM: A Robust Semantic SLAM in Dynamic Environments Based on RGB-D Sensor.” In: *IEEE Sensors Journal* (2021).

[161] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. “R2D2: Reliable and repeatable detector and descriptor.” In: *NeurIPS*. 2019.

[162] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. "Efficient neighbourhood consensus networks via submanifold sparse convolutions." In: *ECCV*. 2020.

[163] Alberto Rodrıéguez, J Javier Yebes, Pablo F Alcantarilla, Luis M Bergasa, Javier Almazán, and Andrés Cela. "Assisting the visually impaired: obstacle detection and warning system by acoustic feedback." In: *Sensors* (2012).

[164] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation." In: *T-ITS* 19.1 (2018), pp. 263–272.

[165] Eduardo Romera, Luis M. Bergasa, Kailun Yang, Jose M Alvarez, and Rafael Barea. "Bridging the day and night domain gap for semantic segmentation." In: *IV*. 2019.

[166] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation." In: *MICCAI*. 2015.

[167] Jan Roters, Xiaoyi Jiang, and Kai Rothaus. "Recognition of traffic lights in live video streams on mobile devices." In: *TCSVT* (2011).

[168] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. "ORB: An efficient alternative to SIFT or SURF." In: *ICCV*. 2011.

[169] Avishkar Saha, Oscar Mendez Maldonado, Chris Russell, and Richard Bowden. "Translating images into maps." In: *ICRA*. 2022.

[170] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. "Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments." In: *ASSETS*. 2019.

[171] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. "ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding." In: *ICCV*. 2021.

[172] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "MobileNetV2: Inverted residuals and linear bottlenecks." In: *CVPR*. 2018.

[173] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. "From coarse to fine: Robust hierarchical localization at large scale." In: *CVPR*. 2019.

[174] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "SuperGlue: Learning feature matching with graph neural networks." In: *CVPR*. 2020.

[175] Manolis Savva et al. "Habitat: A Platform for Embodied AI Research." In: *ICCV*. 2019.

[176] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. "Efficient RGB-D semantic segmentation for indoor scene analysis." In: *ICRA*. 2021.

[177] Mehran Shakerinava and Siamak Ravanbakhsh. "Equivariant Networks for Pixelized Spheres." In: *ICML*. 2021.

[178] Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, and Di Lin. "RANet: Region Attention Network for Semantic Segmentation." In: *NeurIPS*. 2020.

[179] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. "Efficient attention: Attention with linear complexities." In: *WACV*. 2021.

[180] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. "UrbanLF: A Comprehensive Light Field Dataset for Semantic Segmentation of Urban Scenes." In: *TCSVT* (2022).

[181] Wenjun Shi et al. "Multilevel cross-aware RGBD indoor semantic segmentation for bionic binocular robot." In: *T-MRB* 2.3 (2020), pp. 382–390.

[182] Wenjun Shi et al. "RGB-D Semantic Segmentation and Label-Oriented Voxelgrid Fusion for Accurate 3D Semantic Mapping." In: *TCSVT* 32.1 (2022), pp. 183–197.

[183] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. "ClusterGNN: Cluster-based Coarse-to-Fine Graph Neural Network for Efficient Feature Matching." In: *CVPR*. 2022.

[184] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. “PST900: RGB-thermal calibration, dataset and segmentation network.” In: *ICRA*. 2020.

[185] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor segmentation and support inference from RGBD images.” In: *ECCV*. 2012.

[186] Walter CSS Simões and VF De Lucena. “Blind user wearable audio assistance for indoor navigation based on visual markers and ultrasonic obstacle detection.” In: *2016 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE. 2016, pp. 60–63.

[187] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar. “Self-supervised Feature Learning for Semantic Segmentation of Overhead Imagery.” In: *BMVC*. 2018.

[188] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. “SUN RGB-D: A RGB-D scene understanding benchmark suite.” In: *CVPR*. 2015.

[189] Carsten Stahlschmidt, Sebastian von Camen, Alexandros Gavriilidis, and Anton Kummert. “Descending step classification using time-of-flight sensor data.” In: *IV*. 2015.

[190] Julian Straub et al. “The Replica dataset: A digital replica of indoor spaces.” In: *arXiv preprint arXiv:1906.05797* (2019).

[191] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. “Segmenter: Transformer for Semantic Segmentation.” In: *ICCV*. 2021.

[192] Cheng Sun, Min Sun, and Hwann-Tzong Chen. “HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features.” In: *CVPR*. 2021.

[193] Dongming Sun, Xiao Huang, and Kailun Yang. “A multimodal vision sensor for autonomous driving.” In: *SPIE*. 2019.

[194] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. “LoFTR: Detector-free local feature matching with transformers.” In: *CVPR*. 2021.

[195] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. “Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images.” In: *RA-L* 5.4 (2020), pp. 5558–5565.

[196] Yuxiang Sun, Weixun Zuo, and Ming Liu. “RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes.” In: *RA-L* 4.3 (2019), pp. 2576–2583.

[197] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. “FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion.” In: *T-ASE* (2021).

[198] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. “FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion.” In: *T-ASE* 18.3 (2021), pp. 1000–1011.

[199] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. “ESS: Learning event-based semantic segmentation from still images.” In: *ECCV*. 2022.

[200] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. “Gated-SCNN: Gated shape CNNs for semantic segmentation.” In: *ICCV*. 2019.

[201] Haobin Tan, Chang Chen, Xinyu Luo, Jiaming Zhang, Constantin Seibold, Kailun Yang, and Rainer Stiefelhagen. “Flying guide dog: Walkable path discovery for the visually impaired utilizing drones and transformer-based semantic segmentation.” In: *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2021.

[202] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. “QuadTree Attention for Vision Transformers.” In: *ICLR* (2022).

[203] Keisuke Tateno, Nassir Navab, and Federico Tombari. “Distortion-aware convolutional filters for dense prediction in panoramic images.” In: *ECCV*. 2018.

[204] Zhifeng Teng*, Jiaming Zhang*, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. “360BEV: Panoramic Semantic Mapping for Indoor Bird’s-Eye View.” In: *WACV* (2024).

[205] Shishun Tian, Minghuo Zheng, Wenbin Zou, Xia Li, and Lu Zhang. “Dynamic Crosswalk Scene Understanding for the Visually Impaired.” In: *IEEE Trans. Neural Syst. Rehabil. Eng.* (2021).

[206] Ilya O. Tolstikhin et al. “MLP-mixer: An all-MLP Architecture for Vision.” In: *NeurIPS*. 2021.

[207] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention.” In: 2020.

[208] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. “Learning to Adapt Structured Output Space for Semantic Segmentation.” In: *CVPR*. 2018.

[209] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. “Self-supervised model adaptation for multimodal semantic segmentation.” In: *IJCV* 128.5 (2019), pp. 1239–1285.

[210] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. “Self-supervised model adaptation for multimodal semantic segmentation.” In: *IJCV* (2020).

[211] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. “IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments.” In: *WACV*. 2019.

[212] Ashish Vaswani et al. “Attention is all you need.” In: *NeurIPS*. 2017.

[213] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All you Need.” In: *NeurIPS*. 2017.

[214] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. “ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation.” In: *CVPR*. 2019.

[215] Hao Wang, Weining Wang, and Jing Liu. “Temporal memory attention for video semantic segmentation.” In: *ICIP*. 2021.

[216] Hsueh-Cheng Wang, Robert K. Katzschmann, Santani Teng, Brandon Araki, Laura Giarré, and Daniela Rus. “Enabling independent navigation for visually impaired people through a wearable vision-based feedback system.” In: *ICRA*. 2017.

[217] Jingdong Wang et al. “Deep high-resolution representation learning for visual recognition.” In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[218] Jingdong Wang et al. “Deep high-resolution representation learning for visual recognition.” In: *TPAMI* 43.10 (2021), pp. 3349–3364.

[219] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. “Learning feature descriptors using camera pose supervision.” In: *ECCV*. 2020.

[220] Weiyue Wang and Ulrich Neumann. “Depth-aware CNN for RGB-D segmentation.” In: *ECCV*. 2018.

[221] Wenhai Wang et al. “PVTv2: Improved Baselines with Pyramid Vision Transformer.” In: *arXiv preprint arXiv:2106.13797* (2021).

[222] Wenhai Wang et al. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.” In: *ICCV*. 2021.

[223] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. “Non-local neural networks.” In: *CVPR*. 2018.

[224] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. “Multimodal Token Fusion for Vision Transformers.” In: *CVPR*. 2022.

[225] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. “Learning Deep Multimodal Feature Representation with Asymmetric Multi-layer Fusion.” In: *MM*. 2020.

[226] Yu Wang et al. "LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation." In: *ICIP*. 2019.

[227] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogério Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S. Huang, and Honghui Shi. "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation." In: *CVPR*. 2020.

[228] Qing Wang*, Jiaming Zhang*, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. "Matchformer: Interleaving attention in transformers for feature matching." In: *ACCV*. 2022.

[229] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. "P2T: Pyramid Pooling Transformer for Scene Understanding." In: *arXiv* (2021).

[230] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. "CGNet: A light-weight context guided network for semantic segmentation." In: *IEEE Trans. Image Process.* (2021).

[231] Tianyi Wu, Sheng Tang, Rui Zhang, and Yongdong Zhang. "CGNet: A Light-Weight Context Guided Network for Semantic Segmentation." In: *TIP* 30 (2021), pp. 1169–1179.

[232] Zhongwei Wu, Zhuyun Zhou, Guillaume Allibert, Christophe Stolz, Cédric Demonceaux, and Chao Ma. "Transformer Fusion for Indoor RGB-D Semantic Segmentation." In: *CVIU* (2022).

[233] Zongwei Wu, Guillaume Allibert, Christophe Stolz, and Cédric Demonceaux. "Depth-Adapted CNN for RGB-D cameras." In: *ACCV*. 2020.

[234] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. "Gibson env: Real-world perception for embodied agents." In: *CVPR*. 2018.

[235] Kaite Xiang, Kailun Yang, and Kaiwei Wang. "Polarization-driven semantic segmentation via efficient attention-bridged fusion." In: *Opt. Express* (2021).

[236] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. "Unified perceptual parsing for scene understanding." In: *ECCV*. 2018.

[237] Enze Xie et al. "Segmenting transparent object in the wild with transformer." In: *IJCAI*. 2021.

[238] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." In: *NeurIPS*. 2021.

[239] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." In: *NeurIPS*. 2021.

[240] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. "Segmenting transparent objects in the wild." In: *ECCV*. 2020.

[241] Yajie Xing, Jingbo Wang, and Gang Zeng. "Malleable 2.5D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing." In: *ECCV*. 2020.

[242] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. "PAD-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing." In: *CVPR*. 2018.

[243] Jiangtao Xu, Kaige Lu, and Han Wang. "Attention fusion network for multi-spectral semantic segmentation." In: *PRL* 146 (2021), pp. 179–184.

[244] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. "Semantic segmentation of panoramic images using a synthetic dataset." In: *SPIE*. 2019.

[245] Ran Yan, Kailun Yang, and Kaiwei Wang. "NLFNet: Non-local Fusion Towards Generalized Multimodal Semantic Segmentation across RGB-depth, Polarization, and Thermal Images." In: *ROBIO*. 2021.

[246] Kailun Yang, Luis M Bergasa, Eduardo Romera, Ruiqi Cheng, Tianxue Chen, and Kaiwei Wang. "Unifying terrain awareness through real-time semantic segmentation." In: *IV*. 2018.

[247] Kailun Yang, Ruiqi Cheng, Luis M Bergasa, Eduardo Romera, Kaiwei Wang, and Ningbo Long. "Intersection perception through real-time semantic segmentation to assist navigation of visually impaired pedestrians." In: *ROBIO*. 2018.

[248] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, Xiao Huang, Dongming Sun, and Kaiwei Wang. "Can we PASS beyond the field of view? Panoramic annular semantic segmentation for real-world surrounding perception." In: *IV*. 2019.

[249] Kailun Yang, Xinxin Hu, Luis M. Bergasa, Eduardo Romera, and Kaiwei Wang. "PASS: Panoramic annular semantic segmentation." In: *IEEE Trans. Intell. Transp. Syst.* (2020).

[250] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. "DS-PASS: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing." In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2020, pp. 457–464.

[251] Kailun Yang, Kaiwei Wang, Xiangdong Zhao, Ruiqi Cheng, Jian Bai, Yongying Yang, and Dong Liu. "IR stereo RealSense: Decreasing minimum range of navigational assistance for visually impaired individuals." In: *Journal of Ambient Intelligence and Smart Environments* (2017).

[252] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. "Capturing Omni-Range Context for Omnidirectional Segmentation." In: *CVPR*. 2021.

[253] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. "DenseASPP for Semantic Segmentation in Street Scenes." In: *CVPR*. 2018.

[254] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. "Learning to find good correspondences." In: *CVPR*. 2018.

[255] Minghao Yin et al. "Disentangled non-local neural networks." In: *ECCV*. 2020.

[256] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. "Disentangled non-local neural networks." In: *ECCV*. 2020.

[257] Senthil Kumar Yogamani et al. "WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving." In: *ICCV*. 2019.

[258] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. "BiSeNet: Bilateral segmentation network for real-time semantic segmentation." In: *ECCV*. 2018.

[259] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. "DS-SLAM: A semantic visual SLAM towards dynamic environments." In: *IROS*. IEEE. 2018.

[260] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning." In: *CVPR*. 2020.

[261] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. "MetaFormer is actually what you need for vision." In: *CVPR*. 2022.

[262] Yuhui Yuan, Xilin Chen, and Jingdong Wang. "Object-contextual representations for semantic segmentation." In: *ECCV*. 2020.

[263] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. "OCNet: Object Context for Semantic Segmentation." In: *Int. J. Comput. Vis.* (2021).

[264] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto L. Sangiovanni-Vincentelli. "Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation." In: *CVPR*. 2021.

[265] Yuchun Yue, Wujie Zhou, Jingsheng Lei, and Lu Yu. "Two-Stage Cascaded Decoder for Semantic Segmentation of RGB-D Images." In: *SPL* 28 (2021), pp. 1115–1119.

[266] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernández Domıénguez. "WildDash - Creating Hazard-Aware Benchmarks." In: *ECCV*. 2018.

[267] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. "Unifying panoptic segmentation for autonomous driving." In: *CVPR*. 2022.

[268] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, and Ian Reid. "Visual odometry revisited: What should be learnt?" In: *ICRA*. IEEE. 2020.

[269] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. "Orientation-Aware Semantic Segmentation on Icosahedron Spheres." In: *ICCV*. 2019.

[270] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. "DeepPanoContext: Panoramic 3D Scene Understanding with Holistic Scene Context Graph and Relation-based Optimization." In: *ICCV*. 2021.

[271] Fan Zhang et al. "ACFNet: Attentional Class Feature Network for Semantic Segmentation." In: *ICCV*. 2019.

[272] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. "Non-Local Aggregation for RGB-D Semantic Segmentation." In: *SPL* 28 (2021), pp. 658–662.

[273] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. "Non-local aggregation for RGB-D semantic segmentation." In: *SPL* (2021).

[274] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. "Context encoding for semantic segmentation." In: *CVPR*. 2018.

[275] Hang Zhang et al. "ResNeSt: Split-attention networks." In: *CVPRW*. 2022.

[276] He Zhang, Lingqiu Jin, and Cang Ye. "An RGB-D Camera Based Visual Positioning System for Assistive Navigation by a Robotic Navigation Aid." In: *IEEE/CAA Journal of Automatica Sinica* 8.8 (2021), pp. 1389–1400.

[277] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. "Learning two-view correspondences and geometry using order-aware network." In: *ICCV*. 2019.

[278] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." In: *T-ITS* (2023).

[279] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. "Delivering Arbitrary-Modal Semantic Segmentation." In: *CVPR*. 2023.

[280] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. "Transfer beyond the Field of View: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation." In: *T-ITS* 23.7 (2022), pp. 9478–9491.

[281] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. "Trans4Trans: Efficient Transformer for Transparent Object Segmentation to Help Visually Impaired People Navigate in the Real World." In: *ICCVW*. 2021.

[282] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. "Trans4Trans: Efficient Transformer for Transparent Object and Semantic Scene Segmentation in Real-World Navigation Assistance." In: *T-ITS* 23.10 (2022), pp. 19173–19186.

[283] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. "Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation." In: *CVPR*. 2022.

[284] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. "Behind Every Domain There is a Shift: Adapting Distortion-aware Vision Transformers for Panoramic Semantic Segmentation." In: *arXiv preprint arXiv:2207.11860* (2022).

[285] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. "ISSAFE: Improving Semantic Segmentation in Accidents by Fusing Event-based Data." In: *IROS*. 2021.

[286] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. "Exploring Event-driven Dynamic Context for Accident Scene Segmentation." In: *T-ITS* 23.3 (2022), pp. 2606–2622.

[287] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. “ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation.” In: *CVPR*. 2021.

[288] Xiaoya Zhang, Shumin Zhang, Zhen Cui, Zechao Li, Jin Xie, and Jian Yang. “Tube-embedded Transformer for Pixel Prediction.” In: *TMM* 25 (2023), pp. 2503–2514.

[289] Yan Zhang, Ziang Li, Haole Guo, Luyao Wang, Qihe Chen, Wenjie Jiang, Mingming Fan, Guyue Zhou, and Jiangtao Gong. ““I am the follower, also the boss”: Exploring Different Levels of Autonomy and Machine Forms of Guiding Robots for the Visually Impaired.” In: *CHI*. 2023.

[290] Yingzhi Zhang, Haoye Chen, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. “Perception Framework through Real-Time Semantic Segmentation and Scene Recognition on a Wearable System for the Visually Impaired.” In: *RCAR*. 2021.

[291] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. “Pattern-affinitive propagation across depth, surface normal and semantic segmentation.” In: *CVPR*. 2019.

[292] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. “ICNet for real-time semantic segmentation on high-resolution images.” In: *ECCV*. 2018.

[293] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. “Pyramid Scene Parsing Network.” In: *CVPR*. 2017.

[294] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. “Pyramid scene parsing network.” In: *CVPR*. 2017.

[295] Junwei Zheng, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. “MateRobot: Material Recognition in Wearable Robotics for People with Visual Impairments.” In: *arXiv preprint arXiv:2302.14595* (2023).

[296] Sixiao Zheng et al. “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers.” In: *CVPR*. 2021.

[297] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. “Complementary Bi-directional Feature Compression for Indoor 360°Semantic Segmentation with Self-distillation.” In: *arXiv preprint arXiv:2207.02437* (2022).

[298] Guo Zhong and Chi-Man Pun. “Subspace clustering by simultaneously feature selection and similarity learning.” In: *Knowledge-Based Systems* (2020).

[299] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. “Scene Parsing through ADE20K Dataset.” In: *CVPR*. 2017.

[300] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. “Understanding the robustness in vision transformers.” In: *ICML*. 2022.

[301] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. “CANet: Co-attention Network for RGB-D Semantic Segmentation.” In: *PR* 124 (2022), p. 108468.

[302] Heng Zhou, Chunna Tian, Zhenxi Zhang, Qizheng Huo, Yongqiang Xie, and Zhongbo Li. “Multi-spectral Fusion Transformer Network for RGB-thermal Urban Scene Semantic Segmentation.” In: *GRSL* (2022).

[303] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. “Decoupled dynamic filter networks.” In: *CVPR*. 2021.

[304] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. “Patch2Pix: Epipolar-guided pixel-level correspondences.” In: *CVPR*. 2021.

[305] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. “GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation.” In: *TIP* 30 (2021), pp. 7790–7802.

[306] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. “PGDENet: Progressive Guided Fusion and Depth Enhancement Network for RGB-D Indoor Scene Parsing.” In: *TMM* (2022).

[307] Zhili Zhou, Q. M. Jonathan Wu, Shaohua Wan, Wendi Sun, and Xingming Sun. "Integrating SIFT and CNN feature matching for partial-duplicate image detection." In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020).

[308] Jiafan Zhuang, Zilei Wang, and Bingke Wang. "Video semantic segmentation with distortion-aware feature correction." In: *TCSVT* (2021).

[309] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. "Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation." In: *ICCV*. 2021.

[310] Wenbin Zou, Guoguang Hua, Yue Zhuang, and Shishun Tian. "Real-time Passable Area Segmentation with Consumer RGB-D Cameras for the Visually Impaired." In: *TIM* (2023).

[311] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. "Confidence Regularized Self-Training." In: *ICCV*. 2019.