International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02376-9

DustNet++: Deep Learning-Based Visual Regression for Dust Density Estimation

Andreas Michel[1,2] · Martin Weinmann[2] · Jannick Kuester[1] · Faisal AlNasser[3] · Tomas Gomez[4] · Mark Falvey[4] · Rainer Schmitz[4] · Wolfgang Middelmann[1] · Stefan Hinz[2]

Received: 28 May 2024 / Accepted: 4 February 2025

Abstract

Detecting airborne dust in standard RGB images presents significant challenges. Nevertheless, the monitoring of airborne dust holds substantial potential benefits for climate protection, environmentally sustainable construction, scientific research, and various other fields. To develop an efficient and robust algorithm for airborne dust monitoring, several hurdles have to be addressed. Airborne dust can be opaque or translucent, exhibit considerable variation in density, and possess indistinct boundaries. Moreover, distinguishing dust from other atmospheric phenomena, such as fog or clouds, can be particularly challenging. To meet the demand for a high-performing and reliable method for monitoring airborne dust, we introduce DustNet++, a neural network designed for dust density estimation. DustNet++ leverages feature maps from multiple resolution scales and semantic levels through window and grid attention mechanisms to maintain a sparse, globally effective receptive field with linear complexity. To validate our approach, we benchmark the performance of DustNet++ against existing methods from the domains of crowd counting and monocular depth estimation using the Meteodata airborne dust dataset and the URDE binary dust segmentation dataset. Our findings demonstrate that DustNet++ surpasses comparative methodologies in terms of regression and localization capabilities.

Keywords Airborne dust detection · Machine learning · Visual regression · Attention

1 Introduction

Monitoring airborne dust emissions is an essential endeavor due to its profound impact on climate, human health, infrastructure, buildings, and various socio-economic sectors. The source of airborne dust particles can stem from natural phenomena such as strong winds, wildfires, and seismic activities, as well as from human activities. Predominant anthropogenic sources include construction sites, vehicular traffic, and mining operations. While the complete eradication of dust emissions is unattainable, targeted suppression measures can be implemented. These measures might include but are not limited to watering untreated roads, deceler-

Communicated by Ullrich Köthe.

✉ Andreas Michel
andreas.michel@iosb.fraunhofer.de

Martin Weinmann
martin.weinmann@kit.edu

Jannick Kuester
jannick.kuester@iosb.fraunhofer.de

Faisal AlNasser
alnasser@mit.edu

Tomas Gomez
tomas@meteodata.cl

Mark Falvey
falvey@meteodata.cl

Rainer Schmitz
schmitzr@meteodata.cl

Wolfgang Middelmann
wolfgang.middelmann@iosb.fraunhofer.de

Stefan Hinz
stefan.hinz@kit.edu

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Ettlingen, Germany

[2] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany

[3] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, USA

[4] Meteodata, Santiago, Chile

Published online: 24 February 2025

ating vehicles, or curtailing mining activities. However, effective and economical monitoring of dust emissions is imperative to enhance dust mitigation strategies. Conventional instrument-based *in-situ* monitoring equipment does not focus on identifying emissions but on concentrations of particles.

On the contrary, while insightful, advanced remote sensing technologies such as 3D LiDAR scanning are not economically viable on a large scale and often yield noisy, ambiguous data, particularly in complex terrains. Consequently, visual monitoring through camera-based systems stands out as a more feasible option for the detection of airborne dust emissions. Nonetheless, the technique of visual dust density estimation remains significantly underexplored. The scarcity of research in this domain is largely attributable to the intricate challenge of detecting dust in images, a task that is highly ill-posed. Numerous factors contribute to the complexity of dust detection; dust can vary greatly in density and can appear both opaque and translucent. The variation in dust density can be imbalanced, with dense dust from dust storms occurring less frequently in arid regions than the more transparent dust brought about by mild winds, often manifesting sporadically during specific meteorological conditions. Additionally, dust can originate from a multitude of locations and due to various factors. The transparency of dust means that its visual appearance is heavily influenced by environmental conditions, leading to indistinct boundaries in images. Consequently, images depicting dust often appear partially blurred and typically exhibit low spatial contrast. Classical algorithms struggle to capitalize on these partial blur effects because similar effects are also produced by other atmospheric phenomena, such as fog or clouds. Furthermore, while human observers may find it easier to identify dust across a sequence of images, algorithmically harnessing temporal data to improve detection accuracy poses a significant challenge. Moreover, the absence of a consistent color scheme for dust complicates its detection based on visual cues alone. For instance, opaque dust may exhibit a brownish hue during a dust storm or appear black in the aftermath of a mining explosion. Collectively, these characteristics underscore the necessity for a more sophisticated approach to detect and monitor airborne dust emissions effectively.

In the last decade, deep learning has had huge success in various tasks like classification (Krizhevsky et al., 2017), object detection (Ren et al., 2015), neural linguistic processing (Vaswani et al., 2017), and remote sensing (Zhu et al., 2017). However, airborne dust monitoring, obstructed by the aforementioned challenges, is not well researched, and most scientific papers focus on satellite images (Lee et al., 2021), or related tasks like smoke binary segmentation (Yuan et al., 2020). Recently, De Silva et al. published the Unsealed Roads Dust Emissions (URDE) dataset (De Silva et al., 2023) representing a binary dust segmentation dataset. While this can be seen as a first important step towards dust monitoring, we believe that a regression approach could be more beneficial. In contrast to semantic segmentation, which aims to predict labels on a per-pixel basis, the continuous range of dust densities rather suits a regression strategy. Furthermore, the vague boundaries of dust make it challenging to create discrete hard labels.

Accordingly, this work focuses on dust density estimation (see Fig. 1) and thus is most related to DeepDust (Michel et al., 2023). DeepDust estimates dust density maps on single images. In order to detect dust, it exploits multiscale feature maps by utilizing feature pyramid network (FPN) (Lin et al., 2017) structures. In contrast to that work, we also address the fusion of temporal information. Furthermore, we also exploit attention strategies and rely not only on a strictly convolutional method.

In order to validate the effectiveness of our approach, we compare the results achieved to those of visual density estimation techniques originating from other domains, including monocular depth estimation (MDE) and crowd counting. MDE aims at estimating the scene depth on a per-pixel basis, whereas crowd counting aims at approximating the number of people in a given image. Though both tasks differ strongly from dust density estimation, our method is heavily influenced by ideas of both domains.

In summary, the main contributions of this work are as follows:

1. We research the underexplored field of airborne dust density estimation and propose various neural network architectures.
2. Our proposed DustNet architectures combine attention-based and convolutional-based FPN structures to merge local and global features.
3. Our work addresses the fusion of temporal features in the field of dust density estimation.
4. In order to demonstrate the effectiveness of the proposed neural network architectures, we compare the achievements by our novel techniques with those of methods originating from the crowd counting and MDE domains on the Meteodata dust dataset.

This journal paper is an extended version of the conference paper (Michel et al., 2024). In particular, we offer the following contributions over the conference version:

5. We present a novel model architecture, termed DustNet++, which is characterized by a more streamlined design, while simultaneously surpassing the performance of its predecessor presented in the initial conference version.
6. We introduce an innovative cross-multi-axis feature pyramid network that facilitates interactions across various

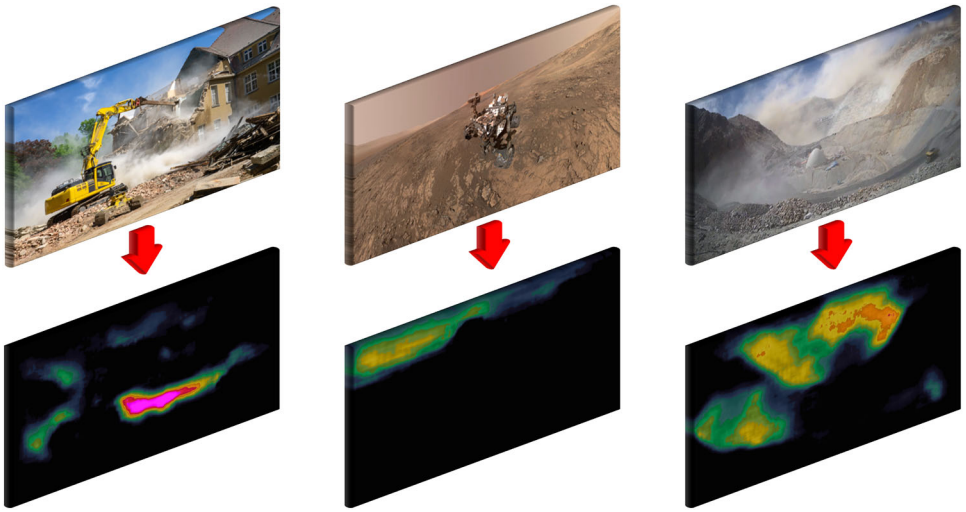

Fig. 1 *DustNet objective.* Our method aims to estimate the level of dust in an RGB image or sequence. The images on the left and right display a construction site (gabort@AdobeStock, 2023) and an opencast mine, respectively. The middle image, on the other hand, shows a scene from the surface of Mars (NASA, 2023). For each image, our DustNet++ model predicts a corresponding dust density map. It is important to note that the goal of our model is not to achieve physically accurate dust estimation, but rather to mimic a human observer who describes the dust levels in a given scene

resolutions and semantic levels of feature maps, while concurrently maintaining both global and local interactions within a feature map.

7. We investigate the poor performance of transformer backbones on the Meteodata dust dataset and conduct a more extensive study on the same dataset.
8. To evaluate the generalization capability of our methodology, we provide a quantitative analysis utilizing the URDE dataset.

2 Related Work

In this section, we briefly summarize related work with a focus on Vision Transformers, crowd counting, and MDE.

2.1 Vision Transformer

Vaswani et al. (2017) introduced the transformative transformer architecture in 2017, which redefined the landscape of natural language processing (NLP). Previously dominated by long short-term memory (Hochreiter & Schmidhuber, 1997) and gated recurrent (Chung et al., 2014) neural networks, the transformer architecture emerged as the new baseline for state-of-the-art methodologies in NLP. Central to this architecture is the multiheaded self-attention (MSA) and cross-attention (MCA), which employ multiple sequences of scaled dot-product attention to facilitate self-attention and cross-attention mechanisms within the neural network.

Building on the enormous success of transformers in NLP, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) adapted the transformer encoder for use in the vision domain. ViT partitions an image into multiple patches and applies global self-attention to analyze these segments. Unlike convolutional neural networks (CNNs), ViT exploits pooling operations, allowing for a more holistic combination of low-level features across the entire image. Although ViT exhibits lower inductive bias, enhancing its scalability (Tay et al., 2022), it also necessitates extended training durations. However, for large images, the computational demands of global self-attention, which scale quadratically, render the basic ViT model less effective.

To address these scalability issues, particularly with large images, Liu et al. presented the Swin Transformer (Swin) (Liu et al., 2021), which introduces a hierarchical structure with shifted windows that effectively reduces computational complexity to a linear scale. This adaptation enhances the feasibility of the transformer for diverse computer vision (CV) applications. The enhanced version of Swin (Liu et al., 2022a) further refines this model by substituting scaled dot product attention with cosine attention, which offers improved performance at larger image size. Similarly, the Multi-axis Vision Transformer (MaxViT) (Tu et al., 2022) merges window and grid attention mechanisms to maintain a sparsely connected yet globally effective receptive field (ERF) with linear complexity, further advancing transformer-based approaches in CV.

Expanding the utility of the Swin Transformer for temporal data, the video Swin Transformer (Liu et al., 2022b) modifies the original architecture to link patches spatially and temporally. Figure 2 emphasizes the distinctions between the global, window, and grid self-attention mechanisms, highlighting their unique attributes and applications.

2.2 Crowd Counting

Density estimation methods have been used successfully in crowd counting (Zhang et al., 2016; Sam et al., 2017; Li et al., 2018; Liu et al., 2019; Luo et al., 2020). The objective of crowd counting is to predict a coarse density map of the relevant target objects, e.g. usually people. The ground truth is generated by smoothing center points with a multi-

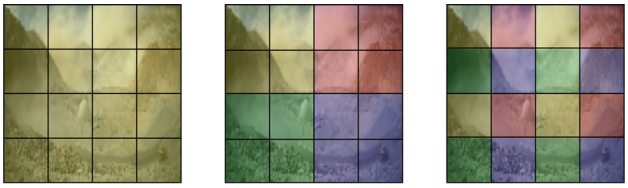

Fig. 2 *Self-attention schemes.* The fundamental concept of Vision Transformers (ViT) (Dosovitskiy et al., 2020) is to divide an image into several patches, convert them into tokens through linear embedding, and then compute the self-attention among them. Originally, ViT applies global attention, where each token can attend to all other elements across the image. However, this approach is not computationally feasible for large images because the computational complexity increases quadratically with the image size. Window attention (Liu et al., 2021), on the other hand, calculates self-attention only for a given block marked by color, reducing the computational complexity to a linear scale, but decreasing the effective receptive field (ERF) significantly. Grid attention (Tu et al., 2022), in contrast, applies a globally sparse uniform grid for a wider ERF at the expense of worse feature flow between neighboring tokens. Similar to window attention, it has linear computational complexity. The grid for the self-attention calculation is denoted by the same color

dimensional Gaussian distribution. Recent approaches are focused on increasing the spatial invariance (Luo et al., 2020) or dealing with noise in the density maps (Cheng et al., 2022). Most works are designed for individual images, but Avvenuti et al. (2022) take advantage of the temporal correlation between consecutive frames in order to reduce localization and count error.

2.3 Monocular Depth Estimation

The first CNN-based method for monocular depth estimation was presented by Eigen et al. (2014). They utilized global and local information in order to predict depth images from a single image. Further improvements of the pure CNN approaches focus on Laplacian pyramids (Song et al., 2021), multi-scale convolutional fusion (Wang et al., 2020), structural information (Lee et al., 2022) or the exploitation of coplanar pixels (Patil et al., 2022) to improve the predicted depth. Recently, hybrids between CNN and Vision Transformer (Dosovitskiy et al., 2020) based architectures improved the depth estimation process. Ranftl et al. (2021) proposed to pass features from a CNN-based ResNet (He et al., 2016) extractor to a Vision Transformer to capture global information. Furthermore, the NeWCRFs (Yuan et al., 2022) approach utilizes window-based Vision Transformers (Liu et al., 2021). Fu et al. (2018) introduced the depth prediction task as a classification–regression problem. Hereby, the classification part consists of predicting discretized bin centers, and the regression part utilizes the bin centers to produce high-quality depth maps. This approach was improved by Bhat et al. (2021) by predicting adaptive bins. The PixelFormer architecture (Agarwal & Arora, 2023) combines transformer architectures with the bin center approach and adds skip connections modules to improve the feature flow between different encoder feature levels.

3 Methodology

In the following, we introduce our proposed DustNet++ architecture illustrated in Fig. 3. Our goal is to simplify the architecture of DustNet (Michel et al., 2024) while still facilitating the interaction between local and global features for high-resolution images. In order to achieve this goal, inspired by MaxViT (Tu et al., 2022), we propose a Cross multi-axis Feature Pyramid Network (CmaxFPN). The basic blocks are a backbone, the CmaxFPN, a feature map matching module and the backend.

3.1 From MaxViT to CmaxFPN

MaxViT (Tu et al., 2022) represents a scalable and efficient approach for computer vision applications by utilizing windowed local and dilated global attention. Let $X \in \mathbb{R}^{H \times W \times C}$ be an input feature map, which is processed by a stem consisting of a convolutional layer and multiple stages of sequences of MaxViT blocks. The basic MaxViT block consists of three modules: an MBConv (Howard et al., 2017) and a pair of self-attention blocks with each respectively using window and grid attention.

MBConv. The MBConv without downsampling can be described as follows:

$$X \leftarrow X + \mathrm{Conv}_{\mathrm{sh}}\left(\mathrm{SE}\left(\mathrm{DWConv}\left(\mathrm{Conv}_{\mathrm{ex}}\left(\mathrm{Norm}(X)\right)\right)\right)\right), \tag{1}$$

where Norm represents Batch Normalization (BatchNorm) (Ioffe & Szegedy, 2015), $\mathrm{Conv}_{\mathrm{ex}}$ denotes the expansion convolution layer in context of the numbers of channels with a kernel of 1×1, DWConv denotes a depthwise convolution layer with a kernel of 3×3, SE denotes the Squeeze Excitation Layer (Hu et al., 2018), and $\mathrm{Conv}_{\mathrm{sh}}$ denotes the corresponding shrinking convolution layer with a kernel of 1×1. Each convolution layer is followed by BatchNorm and the Gaussian Error Linear Unit (GELU) activation function (Hendrycks & Gimpel, 2016). For the first MBConv Block in every stage, the depthwise convolution layer has a stride of two and the skip connection is replaced with a 2D pooling layer.

Multi-Axis Attention. Multi-Axis Attention is based on relative attention (Dai et al., 2021; Shaw et al., 2018). Relative attention adds a relative positional bias B to vanilla self-attention.

B is a learned static location-aware matrix and influences the adaptive attention outputs. Consequently, relative atten-

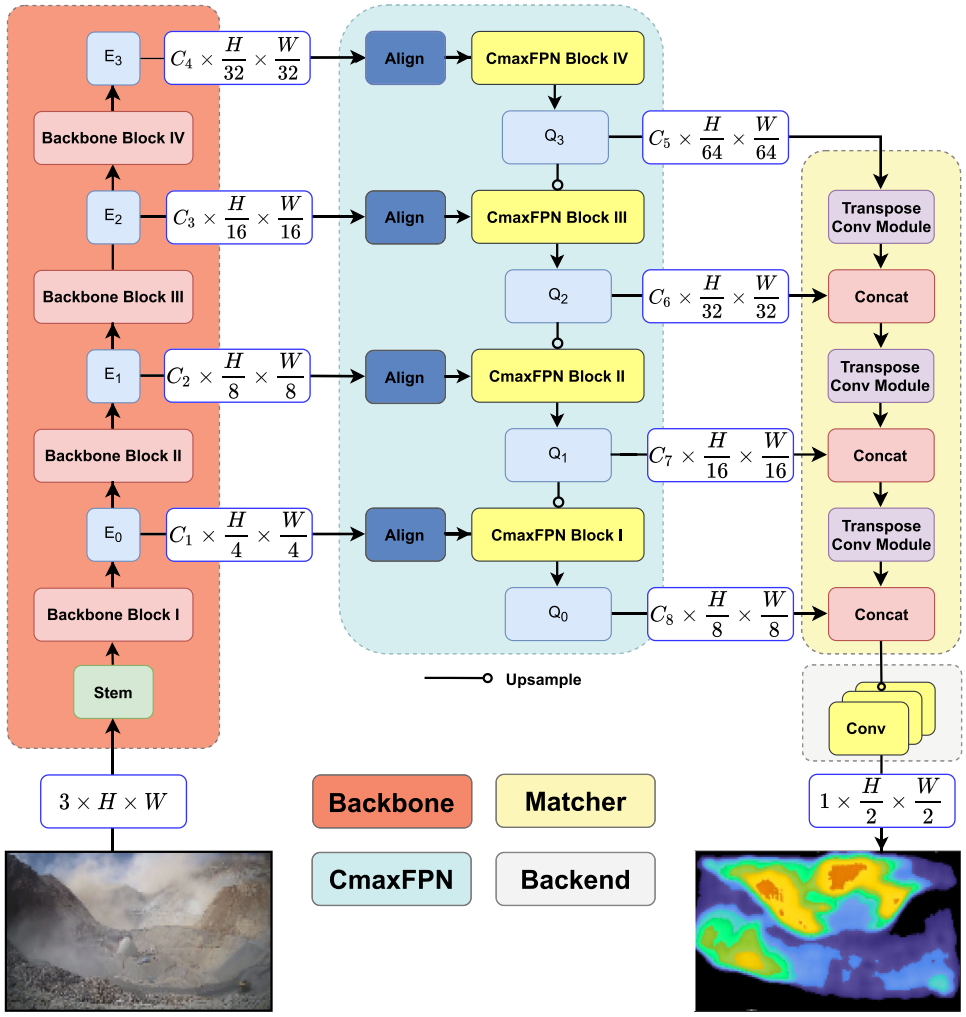

Fig. 3 *Architecture of DustNet++*. The basic components of DustNet++ include the backbone, the CmaxFPN, the matcher, and the backend. These elements, except the Pyramid Pooling Module branch and its adaptive convolution layers, maintain a structural similarity to DustNet. The removal of the latter components has simplified the overall architecture. To facilitate global interactions, the AFPN has been substituted with CmaxFPN. Unlike AFPN, CmaxFPN employs cross-window attention and additionally integrates sparse grid cross-attention, thereby expanding the effective receptive field

tion offers several advantageous features for 2D vision tasks such as input-adaptivity, translation equivariance, and global interactions. In the case of a single head, relative attention is formulated as

$$\text{RelAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{2}$$

where $Q, K, V \in \mathbb{R}^{H\times W\times C}$ are the query, key and value matrices and d is the hidden dimension. In the case of MaxViT, Q, K, V are a linear projection with each weight respectively from the same input vector. Based on the convention for 1D input sequences for Eq. 2, the second last dimension of an input $(..., L, C)$, represents the sequence length or can be defined as the spatial axis. MaxViT does not modify relative attention directly, but modify the input for the spatial axis. In the case of block attention, MaxViT reshapes the input tensor into the shape

$$(H, W, C) \to \left(\frac{H}{P} \times P, \frac{W}{P} \times P, C\right)$$

$$\to \left(\frac{H}{P} \times \frac{W}{P}, P \times P, C\right) \tag{3}$$

with a partition size of P. Therefore, instead of calculating self-attention globally, self-attention is only calculated locally within a partition for window attention. In order to enable sparse global interactions,

$$(H, W, C) \to \left(\frac{H}{G} \times G, \frac{W}{G} \times G, C\right)$$
$$\to \underbrace{\left(G \times G, \frac{H}{G} \times \frac{W}{G}, C\right) \to \left(\frac{H}{G} \times \frac{W}{G}, G \times G, C\right)}_{\text{swapaxes}} \tag{4}$$

The spatial dimensions denoted as H and W are partitioned by the grid size G. This division, followed by transposition and axis swapping, results in the spatial axis representing a fixed uniform grid. The application of self-attention to the reshaped tensor facilitates sparse global interactions. For

both attention mechanisms, an inverse partitioning function is required.

Attention Feature Pyramid Network (AFPN). The AFPN mixes high-resolution features with low-level semantics with low-resolution high-level semantic features. But instead of a traditional FPN like (Lin et al., 2017) utilizing CNNs, we are inspired by Agarwal and Arora (2023) and use four Swin blocks with cross window attention to improve the feature flow between the feature map layers. However, instead of applying scaled dot attention, we utilize cosine attention similar to Liu et al. (2022a) to achieve increased performance in higher resolutions. The key and value matrix inputs K and V are derived from the backbone feature maps, but to reduce computational complexity, we transfer only half of the channels. The query matrix Q is filled by the output of the upsampled stage before. The query matrix with the coarsest resolution originates from the global aggregated features of the Pyramid Pooling Module (PPM) head (Zhao et al., 2017). For the latter, a PPM head like in Yuan et al. (2022) and Agarwal and Arora (2023) can be utilized to aggregate global information of the whole image.

Cross Multi-Axis Feature Pyramid Network (CmaxFPN). The CmaxFPN, a fusion between AFPN and MaxViT (Tu et al., 2022), not only facilitates local–global interactions among the feature maps, but also introduces connections between feature maps across different resolution scales and semantic levels. The reason for the global field of view here in this case is the grid attention introduced by MaxViT, which enables a global field of view without adding another branch like in DustNet.

Initially, the feature maps are uniformly interpolated to a square dimension. The reason hereby is computational efficiency (Tu et al., 2022).

Subsequently, the CmaxFPN for the feature map with the highest semantic level and lowest resolution scale structurally resembles a MaxViT block, where each MBConv (Howard et al., 2017) is succeeded by two pairs of grid and window attention blocks. The motivation behind expanding the original MaxViT block is that the number of layers is similar to the following stages and to avoid architectural complexity. Cross- and self-attention alternate in the window and grid attention pairs in the following stages. Cross-attention is used in order to fuse the features between the the different feature maps:

$$Q_3 \leftarrow \text{CmaxFPN}_{s3}(E_3) \tag{5}$$

$$Q_2 \leftarrow \text{CmaxFPN}_{s2}(E_2, \text{Upsample}(Q_3)) \tag{6}$$

$$Q_1 \leftarrow \text{CmaxFPN}_{s1}(E_1, \text{Upsample}(Q_2)) \tag{7}$$

$$Q_0 \leftarrow \text{CmaxFPN}_{s0}(E_0, \text{Upsample}(Q_1)), \tag{8}$$

where $Q_i, i \in \{0, 1, 2, 3\}$ describes the output for each stage of the CmaxFPN. The upsample operation can be described by the following term:

$$\hat{Q}_i \leftarrow \text{Conv}(\text{PixelShuffle}(Q_i)), \qquad i \in \{0, 1, 2, 3\} \tag{9}$$

Pixel Shuffle (Shi et al., 2016) is employed to upsample the output from the preceding CmaxFPN stage, subsequently aligning it with the channel number of the subsequent feature map E_i, for $i \in \{0, 1, 2, 3\}$. The selection of Pixel Shuffle over alternative methods such as interpolation functions or transpose convolutional layers is predicated on its demonstrated efficacy within the AFPN structure, as introduced in PixelFormer (Agarwal & Arora, 2023) and subsequently employed in DustNet (Michel et al., 2024). Figure 4 delineates the fundamental architecture of a CmaxFPN block, where an MBConv module is succeeded by a pair of window and grid self-attention blocks, mirroring the structure of a MaxViT block. However, the subsequent pair is configured distinctively. Rather than utilizing a single input tensor for query, key, and value matrices, our CmaxFPN employs the upsampled output of the previous stage Q_{i+1} as the query matrix, while the output from the initial multi-axial attention block serves as the key and value matrices. This configuration fosters interactions between features from disparate resolution scales and semantic levels. The implementation of this cross-attention mechanism is a prominent feature of our methodology, distinguishing it from other multi-feature strategies that utilize MaxViT blocks, such as those in MaxViT-UNet (Khan & Khan, 2023), which typically concatenate features across various feature maps. Our approach distinctively leverages the cross-attention scheme to enhance feature integration.

3.2 DustNet++

In Michel et al. (2024), the significance of integrating both global and local features was emphasized. Primarily, the synergistic combination of a broad field of view and high-resolution features enhances DustNet's superior performance. This synergy necessitates a relatively intricate architecture that incorporates an additional branch for high-level semantic features, albeit at a lower spatial resolution. Our proposed subsequent version, DustNet++, strategically eliminates this branch. Nevertheless, in order to maintain a comprehensive field of view, MaxViT blocks are employed in place of the conventional cross Swin blocks. MaxViT blocks exploit grid attention to achieve global feature fusion.

Backbone. Consider an input image $I \in \mathbb{R}^{H \times W \times C}$, where C denotes the number of channels, W the width, and H the height. Analogous to DustNet, the image I is processed through a backbone network, which generates four feature maps $E_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, for $i \in \{0, 1, 2, 3\}$, corresponding to the spatial resolution scales $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of I.

CmaxFPN. Each of the derived feature maps are then aligned by a channel mapper to a fixed number of channels. In contrast to DustNet, the already mentioned high-level seman-

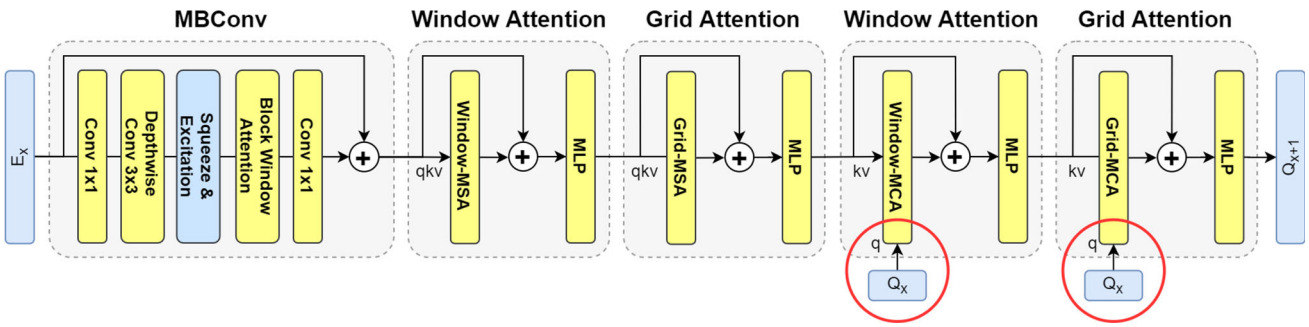

Fig. 4 *CmaxFPN Block.* The CmaxFPN block enhances the MaxViT block by incorporating cross-attention into the basic blocks. Similar to MaxViT, an MBConv is followed by window and grid self-attention layers. In addition, a cross window and cross grid attention layer are included. In this block, the output of the previous feature map is utilized as a query matrix. This facilitates a feature pathway from higher semantic levels with low spatial resolution to features with a lower semantic level and higher spatial resolution

tic branch, which consists of the Pyramid Pooling Module (PPM) and AFPN, has been replaced by the CmaxFPN, whose task is to fuse local and global features (cf. Sect. 3.1).

Matcher. The matcher module remains consistent with that of DustNet, processing the output features from the CmaxFPN. It has an FPN-like architecture (Lin et al., 2017). We upsample the feature maps via a transpose convolution module (TCM). It consists of a 2D TCM with a stride and kernel size of two, followed by batch normalization (Ioffe & Szegedy, 2015) and a SiLU (Elfwing et al., 2018) activation function. As suggested in (Li et al., 2019), we apply only batch normalization without dropout (Srivastava et al., 2014). The coarsest resolution feature map derived from the CmaxFPN is fed to the first TCM block. The following CmaxFPN feature maps are respectively concatenated to the output of the TCM block and processed via the next TCM block.

Backend. The backend consists of N blocks of a sequence of a 2D convolution layer, batch normalization, and SiLU activation functions that predict the dust map. We branch the features into two parallel blocks for each stage and accumulate the outputs. Hereby, we choose a dilation of three for one branch to increase the receptive field. After four stages, a pointwise convolutional layer predicts the dust maps.

3.3 DustNet++ Duo

DustNet++ Duo is heavily inspired by its predecessor, DustNet C. The goal of the design process for the fusion approach is to simplify the architecture, particularly to reduce the number of branches. Most of the design decisions for DustNet++ Duo are motivated by its predecessor, DustNet. Like DustNet C, DustNet++ Duo utilizes two consecutive images that are processed by a shared backbone. But in contrast to DustNet, DustNet++ utilizes grid attention. Grid attention enables DustNet a simple and straightforward way to fuse the images. Before each image's feature maps are sent to CmaxFPN, both images are spatially concatenated along the axis with the smaller dimension. This can be seen in Fig. 5. This configuration enables grid attention to orchestrate the interaction of information between consecutive images. This method presents a more streamlined approach to fusion compared to the techniques previously proposed by DustNet.

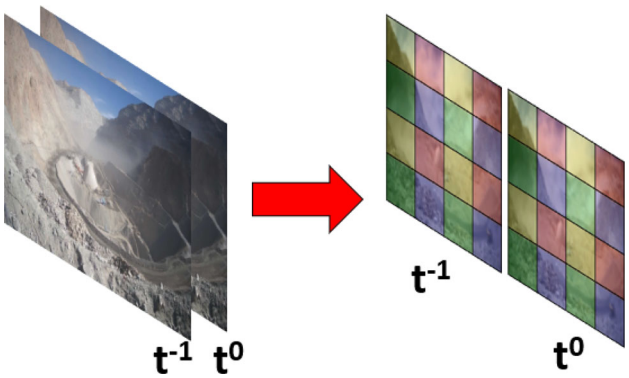

Fig. 5 *Temporal Fusion.* The fundamental premise of the temporal fusion methodology implemented in DustNet++ Duo is to harness the potential of grid attention. This is not solely for the purpose of augmenting the field of view, but also to facilitate the feature flow among temporally sequential images. This is achieved by spatially concatenating the images along the minor dimension after processing through the backbone, and then feeding them into the CmaxFPN

4 Experimental Results

In this section, we outline the implementation specifics of our experiments, expound upon the results obtained, and provide an extensive discussion of our ablation study. Initially, we introduce the datasets employed in our research. Subsequently, we clarify the criteria for benchmark selection and address the intricate details of our experimental implementation. Following this, we engage in both qualitative and quantitative evaluations of the results achieved. Lastly, we elaborate on our ablation study.

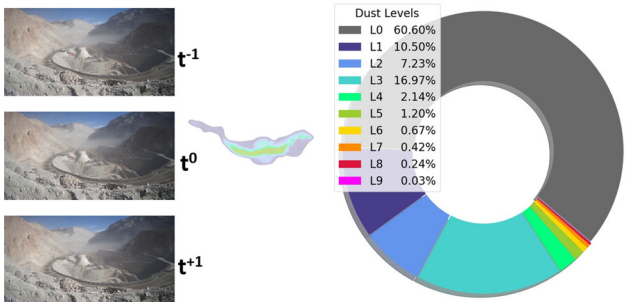

Fig. 6 *Meteodata Dust Dataset.* The Meteodata dust dataset comprises a collection of sequential temporal images accompanied by their respective data distributions. Each data sample includes a triplet of consecutive images captured at varied temporal intervals, wherein the ground truth is associated with the central image. Each image possesses dimensions of 1000×1920 pixels, while the resolution of the ground truth is halved relative to that of the images. Adjacent to the dataset description, a chart visually represents the distribution of labels on a per-pixel basis. Notably, the dataset demonstrates a pronounced imbalance in label distribution, with high-intensity values, potentially indicative of explosive events, being remarkably rare

4.1 Datasets

Our experiments are conducted using two datasets. The focus is set on visual regression for dust density estimation on the temporal Meteodata dust dataset (Michel et al., 2023), whereas the URDE dataset (De Silva et al., 2023) is involved to reason about generalization capabilities of involved approaches.

Meteodata Dust Dataset. This augmented dataset encompasses a diversity of scenes from opencast mines featuring a broad spectrum of dust levels. It includes 2298 RGB image triplets, each with an image size of 1000×1920 pixels. Each triplet comprises three successive images, with the ground truth dust density map corresponding to the middle image. The mean temporal interval between consecutive images is approximately ten seconds. The dust density values in the ground truth are quantified using an 8-bit unsigned integer datatype, with pixel values directly proportional to dust density levels. The ground truth represents only a quarter of the image's total pixels. This is primarily because accurately labeling dust boundaries is challenging, and pixel-perfect labeling may not be practical. Given that the ground truth cannot be pinpointed more precisely, and that a higher resolution would increase computational demands, the decision was made to use a reduced resolution for the ground truth in the dataset.

The dataset is partitioned into training, validation, and testing subsets containing 1906, 144, and 248 image triplets, respectively. The primary challenges associated with this dataset include the computational demand due to size of large images, the difficulty in accurately estimating varied dust levels due to significant variance in dust concentrations, and the pronounced imbalance in the frequency of various dust levels. Figure 6 outlines the ground truth data imbalance. At the time of publication, the Meteodata dust dataset is not publicly available.

URDE Dataset. To assess our methodology's generalization capability, we conducted quantitative and qualitative evaluations on the public available URDE dataset (De Silva et al., 2023). The URDE dataset consists of 7000 images depicting unsealed road segments with ten distinct types of road surface materials under various conditions, designed specifically for the assessment of vehicle-induced road dust emissions. Images captured using a direct single-lens reflex camera were subsequently downsampled to an image size of 1024×1024 pixels. The ground truth data, which required over 1500 h of manual annotation, underpins the dataset's reliability. From the total, 800 images were selected to form the RandomDataset_897 dataset; this subset was divided into 800 images for training and 97 for validation purposes. The selection criteria aimed to minimize the presence of visually similar consecutive images within the samples and to ensure a substantial variation in dust patterns across the dataset.

The original training dataset from URDE RandomDataset_897 was employed solely to evaluate the generalization ability for testing purposes. It is important to note that no training was performed using URDE. The dataset was only used to assess the generalization capacity of the proposed method.

4.2 Evaluation Metrics

We focus on evaluating the localization and regression capabilities of our proposed models. To assess the localization performance, we modify the pixel values, setting all those below 30 to zero and converting the remaining pixels to one. This binary mapping allows us to utilize standard classification metrics such as Accuracy (Acc), Precision (Pre), and Recall (Rec) for performance evaluation. Additionally, to address the challenges posed by imbalanced data distributions commonly encountered in semantic segmentation, we employ the Intersection-over-Union (IoU) metric, which provides a robust measure of overlap between predicted and actual classifications. We use the same binary scheme as for the other classification metrics. Additionally, we utilize the Dice coefficient.

For the validation of regression quality, we utilize conventional metrics including the mean absolute error (MAE) and mean squared error (MSE). These metrics provide insights into the average magnitude of errors in the predictions. Moreover, to ensure a fair evaluation of performance across the tail values in imbalanced datasets, we incorporate the concept of balanced metrics. This approach adjusts the evaluation criteria to give proportional consideration to less frequent, yet significant, data points, thereby providing a more com-

prehensive assessment of model performance across diverse dataset characteristics (Brodersen et al., 2010). Therefore, we bin our data into four bins:

$\mathbf{0-29}$:	Zero dust density bin	(ZB)
$\mathbf{30-99}$:	Low dust density bin	(LB)
$\mathbf{100-169}$:	Medium dust density bin	(MB)
$\mathbf{170-255}$:	High dust density bin	(HB)

For each bin, we calculate the MAE and MSE. Following (Brodersen et al., 2010; Ren et al., 2022), we compute the mean across all bins and obtain the average binned mean absolute error (ØB-MAE) and the average binned mean squared error (ØB-MSE).

4.3 Benchmark Selection

Dust density estimation constitutes a relatively nascent niche within the expansive domain of environmental monitoring. A review of existing literature underscores a paucity of research specifically targeting this area, with notable contributions being DeepDust (Michel et al., 2023) and DustNet (Michel et al., 2024). This observed scarcity underscores the necessity for our benchmark to incorporate methodologies from other fields such as pixel-by-pixel visual regression, drawing insights from crowd counting, and Monocular Depth Estimation (MDE). Our selection criteria were confined to methodologies from different domains that concentrate on the analysis of individual images. This decision was informed by preliminary observations indicating that temporal density estimation methods, commonly employed in crowd counting research (Avvenuti et al., 2022), face convergence challenges in scenarios marked by considerable irregular temporal intervals between successive images of large size. In our methodological framework, we juxtapose DeepDust and DustNet with CanNet (Liu et al., 2019), a context-aware, lightweight, fully convolutional network tailored for crowd counting. The rationale for selecting CanNet hinges on its foundational employment of the VGG16 architecture, which provides a basis for assessing the potential applicability of simpler network structures to temporal dust density estimation tasks. Furthermore, our exploration extends to cutting-edge MDE models, specifically NeWCRF (Yuan et al., 2022) and PixelFormer (Agarwal & Arora, 2023), both of which incorporate the Swin Transformer as their backbone. These models are configured with a window size of twelve and are tested in both base and large variants. This selection is aimed at evaluating the feasibility of advanced MDE techniques in addressing the specific challenges associated with temporal dust density estimation.

4.4 Implementation Details

The experimentation was facilitated using an array of four A100 Nvidia GPUs, each equipped with 80 GB of memory. The foundational benchmark (Michel et al., 2023, 2024) employed the L2 loss function alongside the AdamW optimizer (Loshchilov & Hutter, 2017) (with β values of 0.9 and 0.999) and a weight decay parameter set at 10^{-5}. Throughout the training phase, a learning rate of $\alpha = 3 \times 10^{-4}$ was maintained, and the models were subject to a training duration of 50 epochs with a non-trainable (frozen) backbone. Subsequent to this phase, the backbone was activated (unfrozen) and underwent further training for an additional 20 epochs. For all training steps, we utilized Gradient Accumulation with an accumlated batch size of 32. We normalize input images using the standard ImageNet values (Krizhevsky et al., 2017).

Alternative experimental configurations were explored in response to the observed suboptimal performance of transformer-based backbones as reported in Michel et al. (2023, 2024). Initially, checkpoint wrappers (CPWs) were implemented to enhance memory efficiency during the training processes. Activation checkpointing, as described in Chen et al. (2016), Jain et al. (2020), involves the temporary removal of specific layer activations during backpropagation, which are then recomputed in the backward pass. This method significantly reduces memory consumption and may facilitate the use of larger batch sizes. The Fairscale (FairScale authors, 2021) implementation of this checkpoint wrapper was utilized on the backbones, effectively allowing for a doubling of the real batch size in several instances.

Further adjustments were made to the learning strategy, extending the training period to 200 epochs with an active backbone, while reducing the learning rate for the backbone layers to ten percent of the initial rate. The properties of the employed models are presented in Table 1. Additionally, a comparative analysis was conducted between traditional backbones such as ResNet101 (He et al., 2016) and the Swin Large Transformer (Liu et al., 2021) to comprehensively evaluate the methodologies.

All models were trained to utilize the Meteodata dust training dataset, ensuring consistency in the data used across different experimental setups.

4.5 Results on the Meteodata Dust Dataset

Table 2 outlines the comparative effectiveness of various density estimation methods applied to the Meteodata dust dataset. The column name #Img describes the number of images processed simultaneously. The tag CPW stands for Checkpoint Wrapper and reduces the memory usage while training and therefore allows increasing the batch size. More detailed results are provided in "Appendix A".

Table 1 Model settings

Model	Backbone	Params (M)	Time (ms)	MeM (GB)	#Img
CanNet	VGG16	18	28.53	6.4	1
NewCRF	Swin-B	140	80.36	8.0	1
NewCRF	Swin-L	270	114.6	9.6	1
PixelFormer	Swin-B	128	66.7	5.6	1
PixelFormer	Swin-L	258	105.8	7.2	1
DeepDust	CNV2-B	101	149.0	28.8	1
DeepDust	Swin-L	207	205.9	29.6	1
DustNet S	r101	68	67.6	12	1
DustNet S	Swin-L	227	116.4	12	1
DustNet++	r101	82	123.6	8	1
DustNet++	Swin-L	235	181.3	9.6	1
DustNet A	r101	65	66.4	10.4	3
DustNet B	r101	67	131.1	12.0	3
DustNet D	r101	86	128.0	11.2	3
DustNet C	r101	68	107.3	12.8	2
DustNet C	Swin-L	227	203.4	14.4	2
DustNet++	r101	83	143.9	9.6	2
DustNet++	Swin-L	236	261.1	12.0	2

The table below outlines the configuration of the training process. The symbol #Img indicates the number of sequential images inputted into the model. When two images are used, the ground truth corresponds to the second image or in the case of three images to the middle image. The time refers to the model's inference time for a 1000 × 1920 pixel image on a Nvidia A100 GPU, repeated 1000× for a batch size of one. This configuration is also applied to determine the maximum memory usage. The abbreviations are as follows: Swin-B stands for Swin Transformer Base, Swin-L for Swin Transformer Large, CNV2-B for ConvNeXt V2 Base, and r101 for ResNet101

Single-Image Analysis. The single-image results are produced by models which process only one image. Firstly, it is noteworthy that models trained with a checkpoint wrapper yielded superior results for almost all cases. This result emphasizes the importance of batch size for performance. Notably, the Swin Transformer Large variants, which benefited from increased batch sizes, surpassed the performance of the ResNet101 variants. Despite this, both DustNet and DustNet++ significantly outperformed all other methods by a substantial margin. CanNet, which was not retrained, demonstrated early convergence prior to reaching 50 epochs in previous experiments, and its relatively compact size allowed for the use of large batch sizes without the need for CPW. However, CanNet exhibited limited regression capabilities at higher dust levels.

The method least benefited from the new training scheme was NeWCRF, which showed a decrease in regression performance despite a slight improvement in localization for the Swin Transformer Base variant. In contrast, while NeWCRF experienced modest gains, PixelFormer saw significant enhancements in both regression and localization capabilities through the new training scheme. Although DeepDust still surpasses PixelFormer, the gap between the two methodologies has narrowed, possibly due to PixelFormer's adaptive bins.

The most successful approaches overall remain DustNet, closely followed by DustNet++. DustNet++ exceeds DustNet in MAE, closely aligns with its overall performance while offering a simpler architectural design.

Multi-Image Analysis. The bottom part of Table 2 presents the outcomes on the temporal Meteodata dust dataset. Similar to the single-image scenario, the new training scheme significantly improved results, with the Swin Transformer Large variants outperforming the ResNet101 variants. Here in this case, DustNet++ demonstrated superior regression capabilities compared to DustNet C, albeit with a slight reduction in IoU. Incorporating temporal information consistently led to enhanced overall results.

Figure 7 illustrates the performance of DustNet++ equipped with a Swin Large Transformer backbone in two opencast mining scenes, showcasing the practical application and effectiveness of the model in real-world settings.

4.6 Results on the URDE Dust Dataset

In our analysis, we employed models trained on the Meteodata dust dataset to evaluate the generalization capability of our approach using binary segmentation on the URDE dataset. Given the binary nature of the labels in this dataset, our investigation was specifically confined to

Table 2 Meteodata dust dataset results

Model	Backbone	CPW	#Img	MAE	MSE	IoU
CanNet	VGG16	×	1	20.21	855.40	0.648
NewCRF	Swin-B	×	1	20.00	822.68	0.635
NewCRF	Swin-B	✓	1	20.83	887.67	0.643
NewCRF	Swin-L	✓	1	19.27	740.54	0.654
PixelFormer	Swin-B	×	1	21.53	825.50	0.641
PixelFormer	Swin-B	✓	1	17.43	645.54	0.697
PixelFormer	Swin-L	✓	1	16.29	576.64	0.713
DeepDust	CNV2-B	×	1	19.60	749.36	0.687
DeepDust	CNV2-B	✓	1	17.11	561.89	0.702
DeepDust	Swin-L	✓	1	15.24	482.90	0.774
DustNet S	r101	×	1	19.27	705.47	0.660
DustNet S	r101	✓	1	14.10	458.96	0.756
DustNet S	Swin-L	✓	1	12.80	**402.20**	**0.793**
DustNet++	r101	✓	1	13.95	497.72	0.762
DustNet++	Swin-L	✓	1	**12.63**	422.33	0.784
DustNet A	r101	×	3	18.77	701.73	0.654
DustNet B	r101	×	3	26.60	1528.08	0.481
DustNet D	r101	×	3	17.44	639.10	0.677
DustNet C	r101	×	2	16.77	601.49	0.685
DustNet C	r101	✓	2	12.60	413.22	0.782
DustNet C	Swin-L	✓	2	12.52	414.92	**0.786**
DustNet++	r101	✓	2	12.20	427.72	0.766
DustNet++	Swin-L	✓	2	**11.73**	**396.10**	0.753

Comparison of the best-performing density estimation methods on the Meteodata dust dataset. The highlighted values represent the best scores for using either one or more images simultaneously

assessing the models' localization ability on the URDE RandomDataset_897 dataset.

Due to the different sensitivities to dust characteristics between the Meteodata dust dataset and the URDE dataset, our analysis primarily focused on the lower 35% of the prediction values generated by the models. This threshold was determined empirically, utilizing 5% intervals to ensure precision in our assessment. The results derived from the URDE RandomDataset_897 training dataset, which comprises 800 images, are detailed in Table 3. Notably, CanNet exhibits high accuracy and precision, yet it demonstrates low recall alongside suboptimal Dice and IoU scores. This suggests a low true-positive rate within a highly imbalanced dataset, rendering CanNet as the only model that did not achieve satisfactory outcomes on the URDE dataset.

Conversely, PixelFormer and NeWCRF delivered superior Dice and IoU scores, with PixelFormer exhibiting marginally better results comparable to those observed in the Meteodata dust dataset. DeepDust transcends both models in terms of Dice and IoU scores, surpassing DustNet as well. Additionally, the impact of the backbone size on the performance of the DeepDust architecture is minimally evident when compared to other models.

DustNet++ is the top performer overall, though its advantage over DeepDust is not markedly substantial. Contrasting with the findings from the Meteodata dust dataset, the localization ability of DustNet++ surpasses that of DustNet, suggesting a potentially lesser degree of overfitting in DustNet++ relative to DustNet.

Figure 8 visually presents qualitative results of DustNet++ alongside DustNet S, both employing Swin Large Transformer backbones, demonstrating their practical application and effectiveness in handling the URDE dataset. Further results are provided in "Appendix C".

4.7 Ablation Study

The outcomes of the ablation study on DustNet++ are presented in Table 4, wherein all models employ the Swin Large Transformer backbone and are trained with activation checkpoint wrappers for 200 epochs. In instances where the MSA block was removed, it was substituted with an MCA block to maintain a consistent overall parameter count.

Replacing MBConv. The removal of the MBConv mandates the integration of a 2D max-pooling layer at the first block of a stage to preserve comparable resolution. Conse-

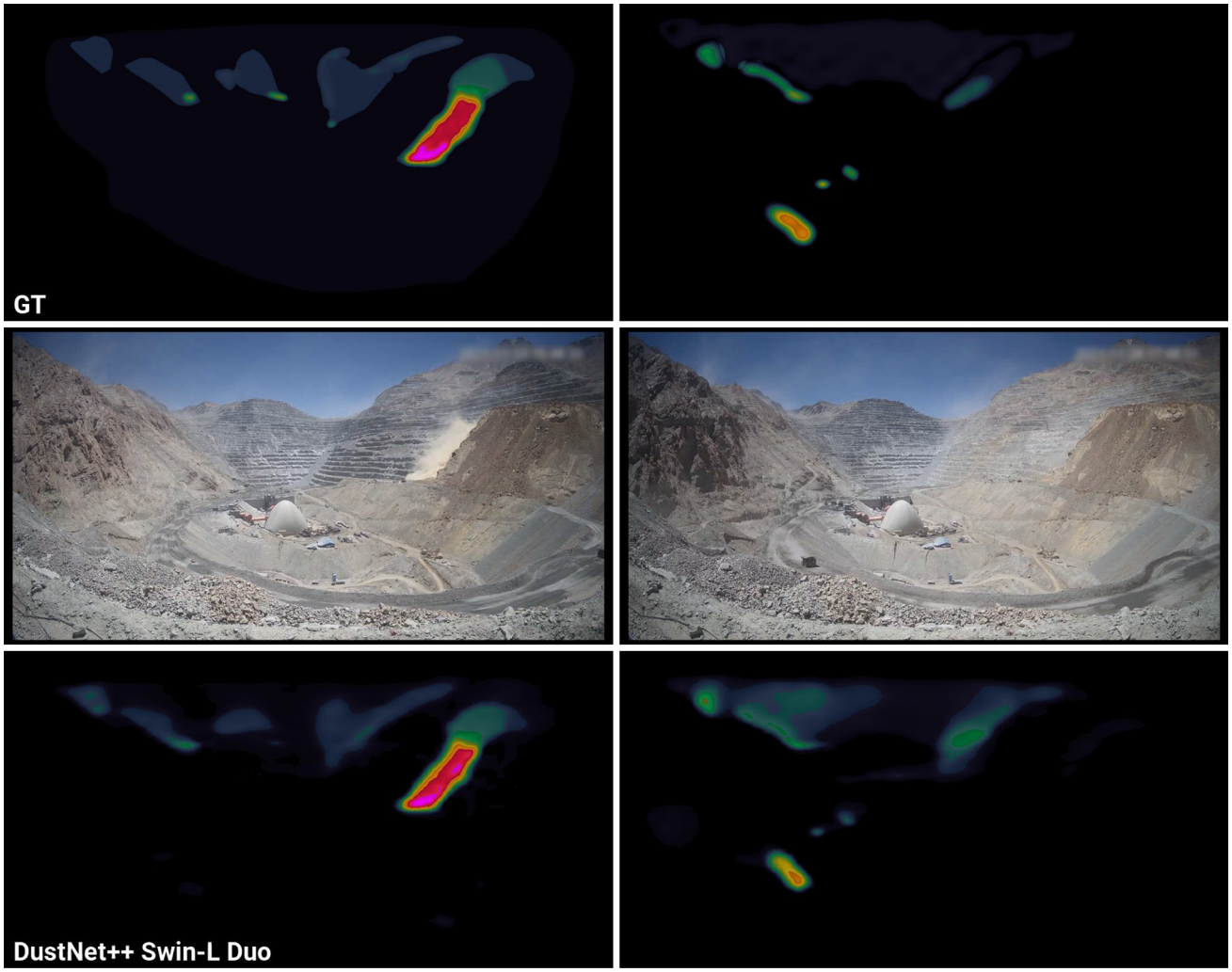

Fig. 7 *Results of DustNet++*. The depicted outcomes encompass two scenes demonstrating the efficacy of DustNet++ Swin-L with two consecutive images as input, which was trained utilizing the Meteodata dust dataset. The scene to the left illustrates the regression proficiency of DustNet++, whereas the scene on the right highlights its localization capabilities

Table 3 URDE Randomdataset_897 dataset

Model	Backbone	CPW	Acc	Pre	Rec	Dice	IoU
CanNet	VGG16	×	**0.972**	0.627	0.506	0.505	0.493
NeWCRF	r101	✓	0.960	0.655	0.712	0.679	0.596
NeWCRF	Swin-L	✓	0.960	0.658	0.723	0.684	0.601
PixelFormer	r101	✓	0.934	0.627	**0.867**	0.679	0.588
PixelFormer	Swin-L	✓	0.947	0.651	0.869	0.708	0.615
DeepDust	CNV2-B	✓	0.951	0.655	0.840	0.708	0.617
DeepDust	Swin-L	✓	0.960	**0.677**	0.798	0.720	0.630
DustNet S	r101	✓	0.942	0.620	0.766	0.660	0.577
DustNet S	Swin-L	✓	0.957	0.663	0.791	0.707	0.618
DustNet++	r101	✓	0.852	0.544	0.714	0.545	0.472
DustNet++	Swin-L	✓	0.957	0.671	0.837	**0.723**	**0.631**

Bold values indicate the best scores
Involved models are trained on the Meteodata dust dataset and then applied with a threshold of 35% on the original Randomdataset_897 training dataset without finetuning

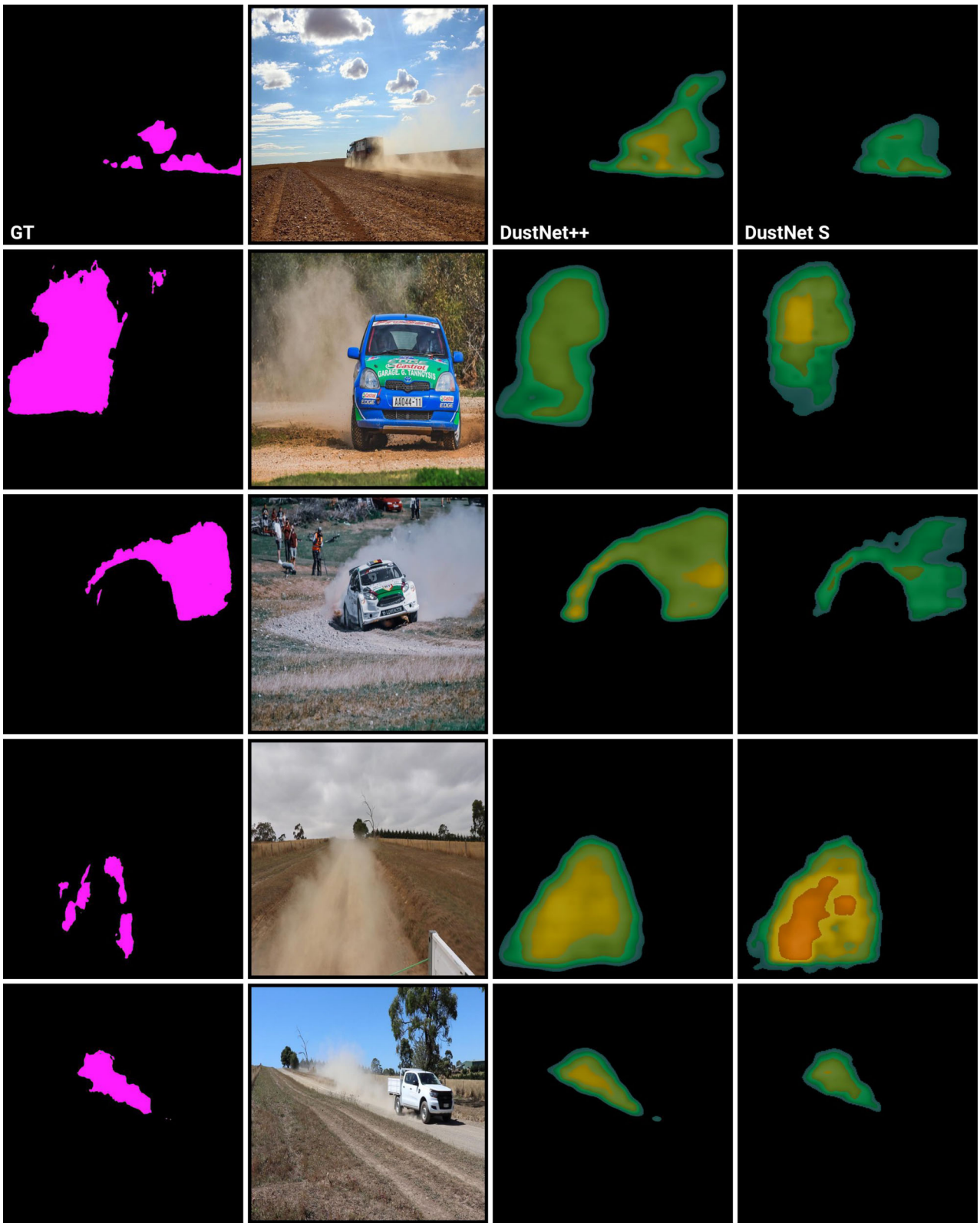

Fig. 8 *Display of the generalization ability of DustNet++*. The shown results of DustNet++ and DustNet S employed with Swin Large Transformer backbones are produced on the URDE validation RandomDataset_897. Hereby, both methods are trained on the Meteodata dust dataset and applied to the URDE dataset without fine-tuning

Table 4 Results of the ablation study on DustNet++

MBConv	✓	×	✓	✓	✓	✓
Cross G-MSA	✓	✓	×	×	✓	✓
Cross W-MSA	✓	✓	×	✓	×	✓
G-MSA	✓	✓	✓	×	✓	×
W-MSA	✓	✓	✓	✓	×	×
MAE	**12.63**	13.74	15.46	15.62	17.94	18.93
MAE-ØB	**22.43**	24.66	22.94	28.35	27.04	36.07
MSE	**422.334**	493.872	512.890	633.247	686.777	813.204
MSE-ØB	**1001.675**	1371.802	1001.675	1503.644	1325.688	2285.558
IoU	**0.784**	0.776	0.761	0.677	0.678	0.634
Dice	**0.879**	0.874	0.864	0.807	0.808	0.775

Bold values indicate the best scores

All models utlize the Swin Large Transformer backbone and are trained with activation checkpoint wrappers for 200 epochs. The elimination of the MBConv necessitates the incorporation of a 2D max-pooling layer at the initial block of a stage and leads to a decrease in the number of parameters. In instances where a multiheaded self-attention (MSA) variant was removed, it was substituted with the remaining MSA block variants to maintain a consistent overall parameter count

quently, this modification results in a reduction in model size. Overall, the omission of the MBConv results in a marginal decline in performance.

Window Attention versus Grid Attention. The exclusion of window attention leads to a more pronounced degradation of overall performance. In contrast, the removal of the grid attention module impairs the model's capacity to discriminate between dust and visually analogous phenomena, such as clouds. Global interactions are critical in reducing the false positive rate associated with similar visual effects, and local interactions are indispensable for precise dust detection. Nevertheless, employing both variants concurrently results in superior outcomes compared to utilizing only one.

MCA versus MSA. Utilizing solely the MCA as opposed to the Vanilla MSA results in a greater decline in performance. This may be attributed to the fact that although the subsequent matcher module facilitates interactions between different feature maps from varying backbone stages, the absence of the Vanilla MSA could potentially lead to disruptions in the feature flow. Analogous to the comparison between window and grid attention, employing both MCA and Vanilla MSA enhances performance.

5 Discussion

5.1 General Discussion

Our studies indicate a significant correlation between batch size and performance. Despite implementing Gradient Accumulation, it was not sufficient to significantly enhance the models' performance. This suggests that performance improvement may not be solely attributed to backpropagation with a larger batch size. Batch normalization could be a determinant factor in this context, particularly as the Meteodata dust dataset's ground truth often presents imprecise boundaries. The impact of metrics in real-world applications, such as in an open-cast mine or construction site, is another essential aspect. An increase in regression ability, gauged by MAE and MSE, may be more pivotal than localization ability, supported by IoU, Accuracy, Recall, and Precision. The ambiguity of dust cloud boundaries makes minor variations in metrics like IoU less crucial. However, differences in regression ability could have a more profound effect. Intense dust events, for instance, those triggered by intentional explosions, are relatively rare but need to be accurately identified in terms of their intensity. All models struggle with such events, particularly those of medium intensity, underlining the importance of a strong regression ability. In the single-image scenario, DustNet++ outperformed DustNet on the original URDE training dataset (cf. Table 3) but slightly underperformed in terms of IoU and MSE on the Meteodata dust validation dataset. Since all methods are trained on the Meteodata dust dataset (cf. Table 2), the generalization ability of the method is more closely examined. In this regard, DustNet++ likely provides superior overall performance. While its regression ability in absolute terms is better, DustNet++ is more vulnerable to outliers than DustNet. For the reasons outlined above, the evaluation can overlook the localization ability closely related to the model, making DustNet++ Duo the evident superior model. In a real-world scenario, DustNet++'s higher inference time compared to DustNet does not pose a significant issue. Given that the processing time is well below one second (cf. Table 1), it is unlikely to exceed the one-second mark even on lower-power GPUs. The lower memory requirement, on the other hand, is a far superior trade-off as it allows for less powerful GPUs or potentially larger images.

5.2 Limitations

Our model demonstrates a tendency to overlook minor dust plumes, which we hypothesize is a trade-off for reducing false positives, ultimately leading to a lower loss during training. This trade-off, however, could be viewed as a negative byproduct of the training process. Additionally, we conjecture that the use of the L2 loss function may exacerbate the mean squared error (MSE) in bins characterized by sparse occurrences. Notably, bins containing rare, high dust values exhibit significantly poorer performance compared to those containing more frequently occurring, lower values.

Moreover, the ground truth for the Meteodata dust dataset is not derived from physical measurements but rather relies on subjective annotations by human annotators. This method of data annotation can introduce a degree of subjectivity and variability in the dataset. Furthermore, the boundaries of dust within this dataset are notably indistinct, leading to heightened levels of ambiguity compared to datasets utilized in classification or object detection tasks. Consequently, there are instances where our model may actually reflect the real dust conditions more accurately than the annotated ground truth.

Given these factors, a superior metric result within our testing framework does not necessarily translate to enhanced performance in practical, real-world applications. Although our model is also evaluated using the URDE binary segmentation dust dataset, a comprehensive comparison between density estimation and semantic segmentation approaches to dust remains crucial to fully understanding the capabilities and limitations of these methodologies in varying contexts.

5.3 Societal Impact

Mineral dust poses significant health risks to humans, serving as a medium for transporting biological components such as bacteria, endotoxins, and fungi through atmospheric pathways (Morman & Plumlee, 2014). Although predominantly originating naturally, mineral dust can frequently result from anthropogenic activities. Notable examples include construction sites and stone-crushing plants, where dust adversely affects workers' health (Leghari et al., 2019). Additionally, eroded agricultural fields can impact nearby rural populations, especially due to the potential for carrying pesticides.

Despite its harmful effects, mineral dust remains comparatively under-regulated relative to industrial pollutants (Morman & Plumlee, 2014). This is largely due to its primary impact on smaller, less-monitored populations, such as rural communities and workers. In the United States, the Environmental Protection Agency (EPA) predominantly monitors densely populated cities, leaving smaller communities undermonitored (U.S. Environmental Protection Agency, 2022). Consequently, the limited monitoring efforts perpetuate a feedback loop that intensifies the issue.

To address this problem, the development of effective detection methods is essential. Enhanced detection and monitoring can help mitigate the health risks associated with mineral dust, ensuring better protection for affected populations. Although the total elimination of dust emissions is not feasible, it is possible to implement targeted mitigation strategies. Such measures could encompass, but are not limited to, the hydration of untreated roads, the reduction of vehicle speeds, and the moderation of mining operations. However, the implementation of effective and cost-efficient monitoring systems is crucial to refine and optimize dust mitigation strategies.

We hope that DustNet++, with its enhanced capabilities for detecting airborne dust, will make a substantial contribution to improving dust monitoring methodologies.

5.4 Future Work

The key step to enhancing the efficiency of dust visual regression methods is to develop a larger and improved dust dataset. A semi-automatic method using DustNet++ could be highly beneficial for this purpose. Integrating architectural modifications, such as a transformer encoder-decoder approach seen in DETR (Carion et al., 2020), could also boost performance. However, these methods generally require a larger dataset for training, highlighting the necessity for superior dust datasets. Additionally, combining visual regression with object detection could enable the identification of both the dust source emitter and the dust it produces.

6 Conclusion

In this paper, we have introduced DustNet++, an advanced neural network designed for dust density estimation. DustNet++ calculates the dust density for each pixel in a given image by effectively leveraging and integrating local, global, and temporal information. This model is capable not only of regressing various dust levels but also of distinguishing between dust and similar visual phenomena such as clouds. Our proposed methodology demonstrates superior performance, outpacing all competing approaches in terms of regression capability on the Meteodata dust dataset and exhibiting enhanced localization ability on the URDE dataset.

Appendix A: Additional Results from Models Trained on the Meteodata Dust Dataset

Figures 9, 10 and 11 display additional results from models trained using the Meteodata dust dataset.

Figure 9 shows various mining events with noticeable differences in the size and density of the dust plumes across the images. CanNet generally performs well in terms of regression ability, but it fails to accurately delineate the boundaries of the dust plumes. Specifically, in the image on the left, the

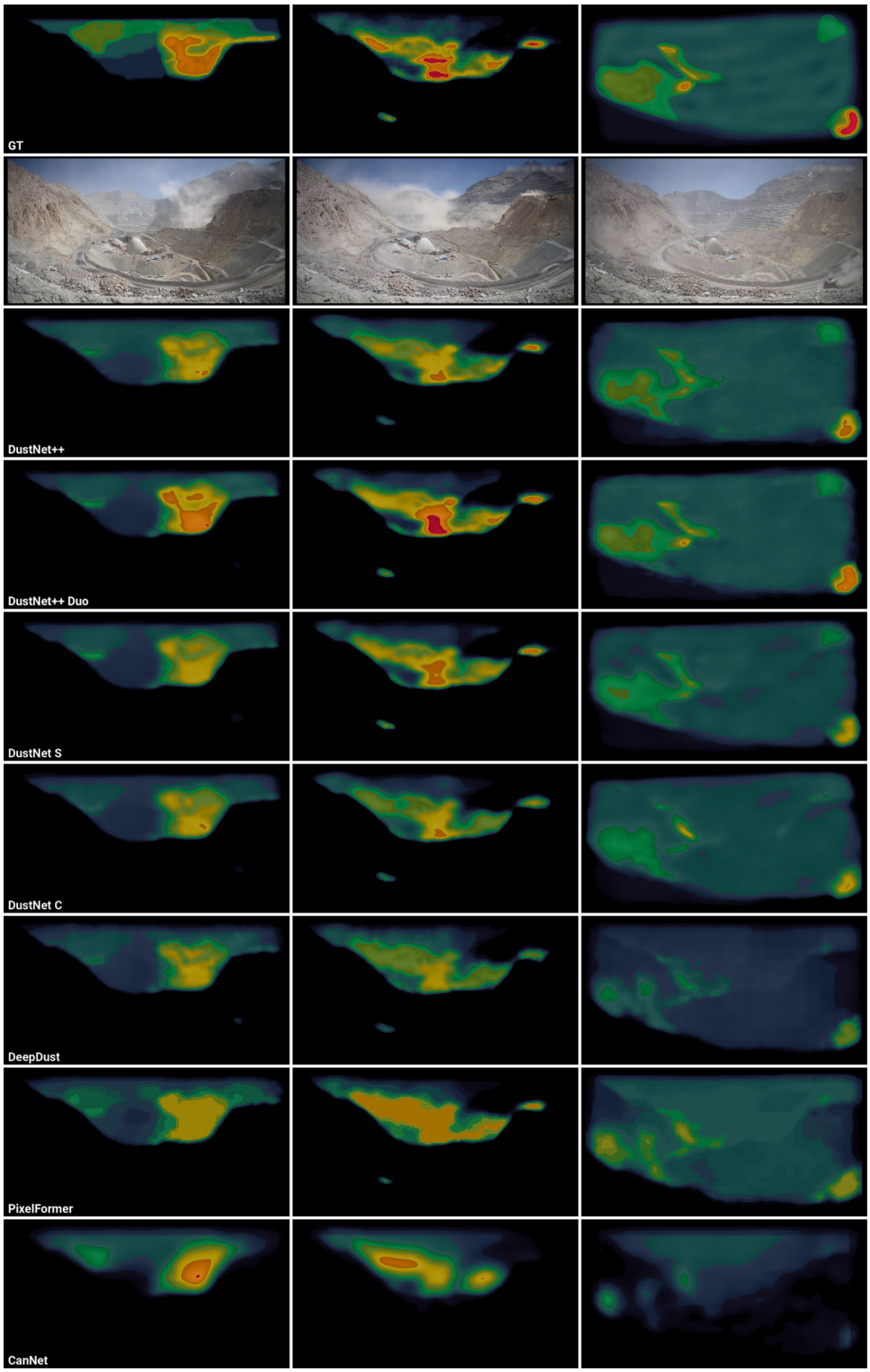

Fig. 9 *Different mining events.* This figure shows various mining events and their associated dust estimation predictions. It is apparent that the considerably smaller CanNet has difficulties with accurate boundary localization, especially with the dust plume in the right images generated by the truck, whereas PixelFormer struggles with effective regression. DeepDust delivers improved outcomes but is readily surpassed by DustNet. Although DustNet S and DustNet++ show similar performance levels, DustNet++ Duo is clearly the best performer in terms of regression ability

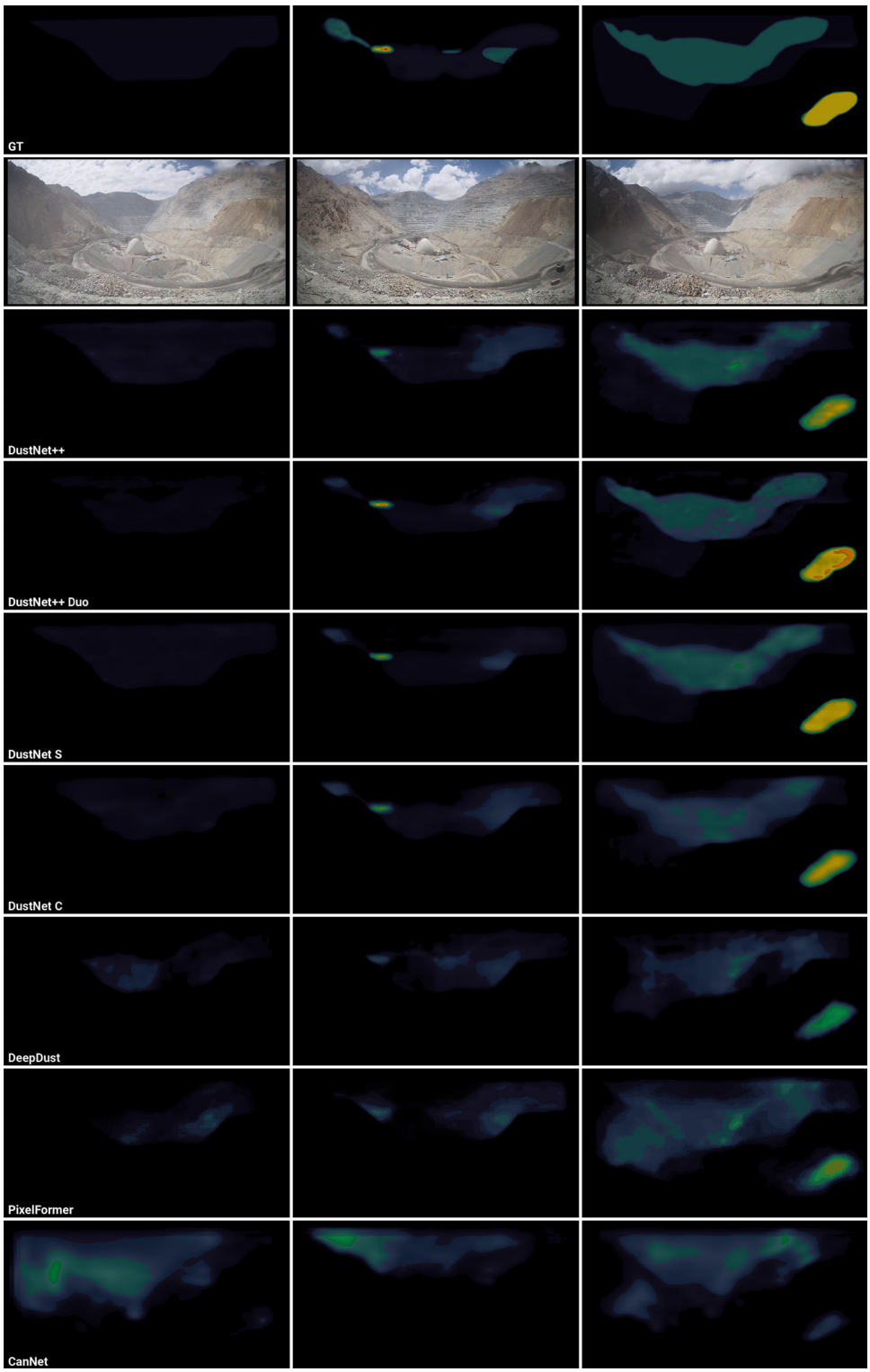

Fig. 10 *Cloudy Scenes*. This figure presents the performance outcomes of models trained using the Meteodata dust dataset in scenarios involving open-cast mine sites with substantial cloud cover. The results demonstrate that DustNet++ adeptly distinguishes between dust particles and cloud formations. In contrast, the rudimentary CanNet model frequently misclassifies clouds as dust. Specifically, in the image on the left, only DustNet++ and DustNet S correctly ascertain the absence of dust, while other models incorrectly detect its presence. In the central image, all models, aside from CanNet, exhibit reasonable performance. In the rightmost image, DustNet S achieves the most precise regression results in the right image, closely followed by DustNet++. The performance of the remaining models is markedly poorer by comparison

dust plume generated by the truck is not tracked correctly, indicating that CanNet's simpler architecture might be insufficient for this task. On the other hand, PixelFormer exhibits superior localization capabilities, although its regression performance is deficient. DeepDust shows enhanced outcomes but is outperformed by DustNet. DustNet S exhibits good localization capabilities, potentially better than DustNet++, but its regression performance is marginally inferior.

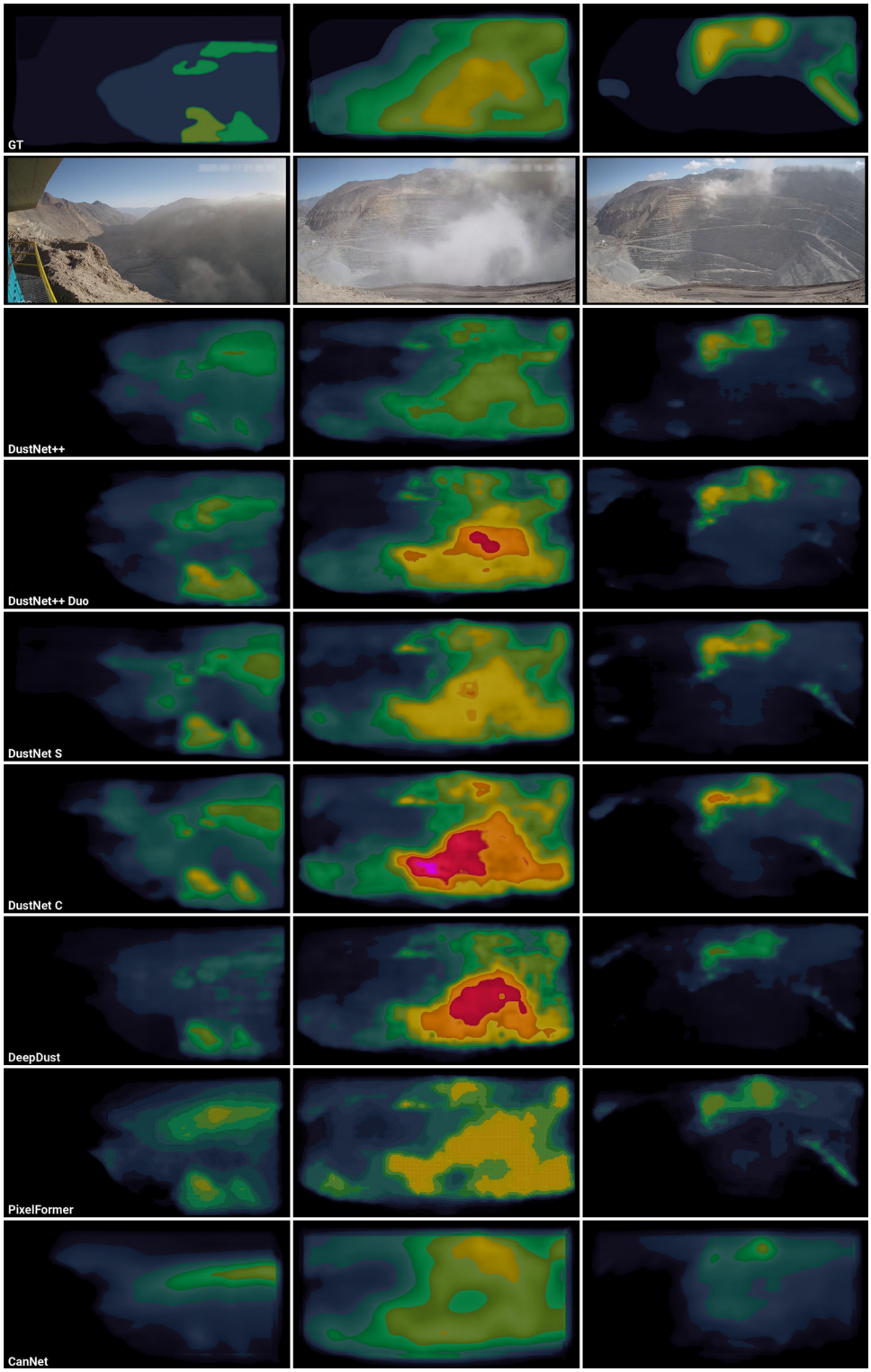

Fig. 11 *Assessment of Model Performance on Unseen Mine Sites.* This figure illustrates the outcomes from models trained using the Meteodata dust dataset on mine sites excluded from the training and validation datasets. The depicted scenes are characterized by their complexity and ambiguity. In the left image, discerning the most effective model is challenging, though it is evident that the basic CanNet model is the least effective. In the central image, the provided ground truth appears to underestimate dust concentrations, suggesting that DustNet++ may offer a more accurate assessment that exceeds these original estimates. Conversely, DeepDust tends to overestimate the dust levels. In the rightmost image, DustNet S outperforms the alternative methodologies

However, DustNet++ Duo significantly outshines the other methods in terms of regression ability in these scenarios.

Figure 10 focuses on scenarios characterized by significant cloud coverage. The data illustrate that DustNet++ effectively differentiates between dust particles and cloud formations. In contrast, the relatively simplistic CanNet model frequently misidentifies clouds as dust. In the left image, only DustNet++ and DustNet S accurately identify the absence of dust, whereas other models erroneously detect dust presence. The central image indicates that all models,

except CanNet, perform adequately. In the right image, DustNet S delivers the most accurate regression results, closely followed by DustNet++. The performance of the other models is notably inferior.

Figure 11 presents the performance of various models on mine sites that were not included in the training and validation datasets of the Meteodata dust project. These scenes are marked by their complexity and ambiguity. In the left image, it is challenging to discern the most effective model, though it is evident that the basic CanNet model performs the least effectively. In the central image, the provided ground truth underestimates dust concentrations, suggesting that DustNet++ might offer a more accurate assessment surpassing these initial estimates. Conversely, DeepDust tends to overestimate the dust levels. In the rightmost image, DustNet S surpasses the performance of the other models.

Appendix B: Detailed Results on the Meteodata Dust Dataset

Tables 5 and 6 supplement Table 2 with more metrics: Accuracy, Precision, Recall, HB-MSE and HB-MAE. The findings remain very comparable to Table 2. In addition, Tables 7 and 8 provide detailed insights into individual MSE and MAE bins.

Table 5 Comparison of the best-performing density estimation methods on the Meteodata dust dataset for single images

Model	Backbone	CPW	#Img	MAE	MSE	ØB-MAE	ØB-MSE	Acc	Pre	Rec	IoU
CanNet	VGG16	×	1	20.21	855.40	38.08	2917.52	0.787	0.798	0.790	0.648
NeWCRF	Swin-B	×	1	20.00	822.68	34.19	2254.98	0.779	0.799	0.783	0.635
NeWCRF	Swin-B	✓	1	20.83	887.67	40.61	3218.86	0.784	0.796	0.787	0.643
NeWCRF	Swin-L	✓	1	19.27	740.54	33.09	2019.21	0.792	0.809	0.796	0.654
PixelFormer	Swin-B	×	1	21.53	825.50	40.68	2864.61	0.783	0.805	0.787	0.641
PixelFormer	Swin-B	✓	1	17.43	645.54	27.41	1424.52	0.822	0.831	0.824	0.697
PixelFormer	Swin-L	✓	1	16.29	576.64	24.44	1141.12	0.833	0.841	0.835	0.713
DeepDust	CNV2-B	×	1	19.60	749.36	30.11	1640.22	0.782	0.796	0.786	0.687
DeepDust	CNV2-B	✓	1	17.11	561.89	28.28	1373.31	0.826	0.839	0.829	0.702
DeepDust	Swin-L	✓	1	15.24	482.90	27.26	1365.84	0.872	0.877	0.874	0.774
DustNet S	r101	×	1	19.27	705.47	31.35	1756.57	0.796	0.810	0.800	0.660
DustNet S	r101	✓	1	14.10	458.96	24.01	1134.22	0.861	0.868	0.864	0.756
DustNet S	Swin-L	✓	1	12.80	**402.20**	**21.19**	**941.10**	**0.885**	**0.888**	**0.886**	**0.793**
DustNet++	r101	✓	1	13.95	497.72	27.09	1395.00	0.865	0.870	0.867	0.762
DustNet++	Swin-L	✓	1	**12.63**	422.33	22.43	1039.23	0.879	0.883	0.881	0.784

Bold values indicate the best scores

Table 6 Comparison of the best-performing density estimation methods on the temporal Meteodata dust dataset

Model	Backbone	CPW	#Img	MAE	MSE	ØB-MAE	ØB-MSE	Acc	Pre	Rec	IoU
DustNet A	r101	×	3	18.77	701.73	32.21	1837.29	0.792	0.811	0.796	0.654
DustNet B	r101	×	3	26.60	1528.08	64.05	6948.25	0.667	0.754	0.677	0.481
DustNet D	r101	×	3	17.44	639.10	31.13	1767.09	0.809	0.832	0.813	0.677
DustNet C	r101	×	2	16.77	601.49	27.29	1361.71	0.814	0.829	0.818	0.685
DustNet C	r101	✓	2	12.60	413.22	20.08	872.33	0.878	0.881	0.879	0.782
DustNet C	Swin-L	✓	2	12.52	414.92	21.16	919.04	**0.880**	**0.885**	**0.882**	**0.786**
DustNet++	r101	✓	2	12.20	427.72	19.89	861.76	0.868	0.875	0.870	0.766
DustNet++	Swin-L	✓	2	**11.73**	**396.10**	**19.14**	**836.83**	0.860	0.870	0.863	0.753

Bold values indicate the best scores

Table 7 Detailed single-image regression results

Model	Backbone	CPW	#Img	HB-MAE	HB-MSE	MB-MAE	MB-MSE	LB-MAE	LB-MSE
CanNet	VGG16	×	1	74.005	7896.82	42.655	2540.28	23.610	856.36
NeWCRF	Swin-B	×	1	59.809	5369.30	40.818	2371.23	25.242	954.47
NeWCRF	Swin-B	✓	1	82.638	8997.13	43.149	2622.07	24.043	877.14
NeWCRF	Swin-L	✓	1	58.703	4833.36	38.883	2085.80	23.655	845.02
PixelFormer	Swin-B	×	1	87.476	8325.33	36.052	1863.56	24.996	892.47
PixelFormer	Swin-B	✓	1	44.217	2990.18	33.335	1629.00	22.052	762.92
PixelFormer	Swin-L	✓	1	37.233	2203.08	30.220	1365.91	21.075	711.97
DeepDust	CNV2-B	×	1	47.280	3340.16	37.039	1979.88	23.933	873.30
DeepDust	CNV2-B	✓	1	46.871	2917.13	35.477	1694.30	20.061	629.20
DeepDust	Swin-L	✓	1	45.639	2982.74	37.135	1797.57	16.777	485.84
DustNet S	r101	×	1	51.202	3778.13	39.452	2152.64	22.345	754.67
DustNet S	r101	✓	1	40.524	2437.82	30.376	1379.73	16.881	514.35
DustNet S	Swin-L	✓	1	**36.064**	**2023.29**	**25.616**	1091.23	**14.631**	**428.98**
DustNet++	r101	✓	1	47.959	3081.32	36.763	1783.07	16.987	507.18
DustNet++	Swin-L	✓	1	41.204	2394.30	25.892	**1075.72**	15.552	463.08

Bold values indicate the best scores

Binned regression results of density estimation methods on the Meteodata dust dataset for single images: The values of the pixels are binned into zero dust (ZB), low dust (LB), medium dust (MB), and high dust (HB) density

Table 8 Detailed multi-image regression results

Model	Backbone	CPW	#Img	HB-MAE	HB-MSE	MB-MAE	MB-MSE	LB-MAE	LB-MSE
DustNet A	r101	×	3	54.683	4076.94	40.610	2195.47	23.166	794.99
DustNet B	r101	×	3	135.613	19389.38	77.805	6698.77	34.851	1553.73
DustNet D	r101	×	3	52.328	3841.00	41.586	2295.11	22.360	742.86
DustNet C	r101	×	2	45.193	2871.32	33.294	1579.19	21.973	738.89
DustNet C	r101	✓	2	33.714	1849.49	23.348	**929.71**	**15.729**	**485.33**
DustNet C	Swin-L	✓	2	38.652	2026.50	**23.013**	942.24	15.777	494.49
DustNet++	r101	✓	2	31.861	**1662.78**	25.369	1049.18	17.278	555.68
DustNet++	Swin-L	✓	2	**31.785**	1676.82	23.262	999.39	15.970	504.80

Bold values indicate the best scores

Binned regression results of density estimation methods on the temporal Meteodata dust dataset: The values of the pixels are binned into zero dust (ZB), low dust (LB), medium dust (MB), and high dust (HB) density

Appendix C: Results on the URDE Validation Dataset

Table 9 depicts the results from the binary dust segmentation URDE RandomDataset_897 validation dataset, consisting of 97 images. Analogous to the experiments conducted on the RandomDataset_897 training dataset, as referenced in Table 3, models were trained on the Meteodata dust dataset and subsequently tested with a threshold set at 35% of the value range to assess their generalization capabilities. Overall, the outcomes demonstrate a marked resemblance across both dataset variants. Furthermore, the localization proficiency of DustNet++ surpasses that of the competing methodologies.

Table 9 URDE Randomdataset_897 validation dataset

Model	Backbone	CPW	Acc	Pre	Rec	Dice	IoU
CanNet	VGG16	×	**0.974**	**0.700**	0.503	0.500	0.490
NeWCRF	r101	✓	0.962	0.662	0.728	0.689	0.605
NeWCRF	Swin-L	✓	0.961	0.655	0.720	0.681	0.599
PixelFormer	r101	✓	0.933	0.625	**0.883**	0.677	0.587
PixelFormer	Swin-L	✓	0.945	0.645	0.879	0.702	0.610
DeepDust	CNV2-B	✓	0.951	0.652	0.845	0.706	0.615
DeepDust	Swin-L	✓	0.961	0.675	0.812	0.722	0.631
DustNet S	r101	✓	0.939	0.613	0.778	0.654	0.572
DustNet S	Swin-L	✓	0.959	0.662	0.786	0.705	0.616
DustNet++	r101	✓	0.851	0.543	0.722	0.543	0.470
DustNet++	Swin-L	✓	0.957	0.669	0.848	**0.724**	**0.632**

Bold values indicate the best scores
Involved models are trained on the Meteodata dust dataset and then applied with a threshold of 35% on the Randomdataset_897 validation dataset

Appendix D: Influence of the Different Dust Levels

DustNet++ utilizing a ResNet101 backbone is meticulously trained on each bin individually and on the entire dataset for 200 epochs, adhering to configurations comparable to those specified in Table 1. The models are trained using the Meteodata dust training dataset and evaluated on the validation dataset, with the outcomes presented in Table 10.

Training specifically on the low dust (LB) values yields the least favorable mean binned mean squared error (MSE).

Table 10 Training on Different Bins

	LB	MB	HB	All
ØB-MAE	77.45	70.53	69.84	**27.09**
ZB-MAE	10.35	0.67	**0.06**	6.63
LB-MAE	24.73	51.29	57.22	**16.99**
MB-MAE	92.55	87.25	115.03	**36.76**
HB-MAE	182.15	142.93	107.06	**47.96**
ØB-MSE	11178.723	8239.257	8212.195	**1394.996**
ZB-MSE	300.873	12.536	**3.010**	208.415
LB-MSE	971.591	3006.320	3642.044	**507.177**
MB-MSE	9635.054	8364.547	13997.481	**1783.074**
HB-MSE	33807.371	21573.625	15206.244	**3081.318**

The pixel values of the Meteodata dust dataset are categorized into distinct dust density bins: zero dust (ZB), low dust (LB), medium dust (MB), and high dust (HB). DustNet++ is systematically trained on each of these bins using a ResNet101 backbone. It is observed that training on the comprehensive dataset generally yields superior outcomes compared to focusing on specific bins, with the exception of the ZB category

While it surpasses the models trained solely on medium (MB) and high dust (HB) values on the LB values, it is outperformed by the model trained on the holistic dataset. Consequently, training to encompass all bins demonstrates enhanced efficacy compared to an exclusive focus on the LB category. A similar pattern is observed with the models trained on MB and HB values. These models excel within their respective bins but are surpassed by the holistic training approach. Thus, adopting a comprehensive training strategy across all bins consistently leads to more robust results than specializing in individual bins.

Appendix E: DustNet

In the following, we describe the DustNet architecture illustrated in Fig. 12. After presenting the submodules, we focus on the different temporal fusion approaches.

Overview. Let $I \in \mathbb{R}^{H \times W \times C}$ be an input image, where C represents the number of channels, W the width, and H the height. The objective of DustNet is to process I to a continuous dust density map $\hat{y}$ with the dimension $\frac{H}{2} \times \frac{W}{2} \times 1$. The image I is fed into a backbone B, which produces four feature maps $E_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, $i \in \{0, 1, 2, 3\}$ with the corresponding spatial resolution scales $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of I. While decreasing in resolution, the information aggregation of the feature maps are ascending. The backbone features are passed to a Pyramid Pooling Module (PPM) (Zhao et al., 2017) and the Attention Feature Pyramid Network (AFPN). Hereby, in order to reduce the computational complexity, only half of the channels of the feature maps are transferred to the AFPN. The PPM head aggregates global information fed into the AFPN and the fuser module. The AFPN mixes the feature maps of different resolutions and information aggregation levels. The processed feature maps are transferred to the matcher module, accumulating the feature maps into one high-resolution map. Then, the high-resolution features are merged with the global information features aggregated from the PPM head in the fuser module. Eventually, the combined

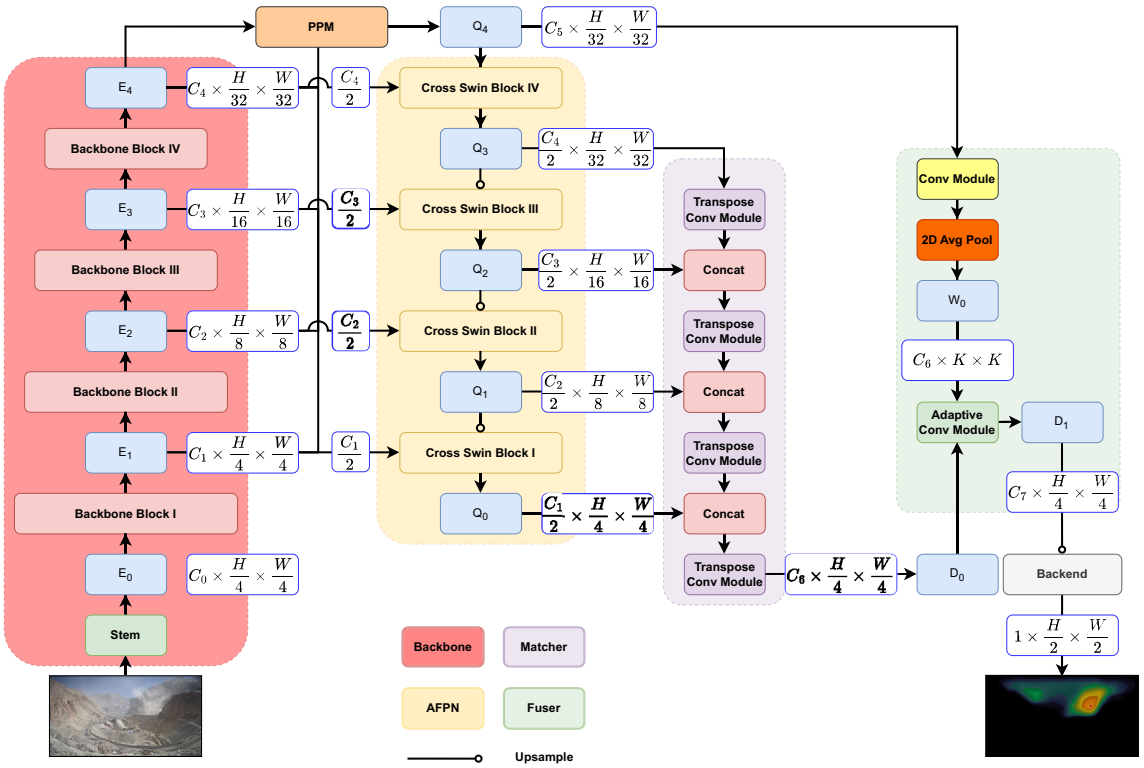

Fig. 12 *Overview of DustNet.* The basic blocks are a backbone, the AFPN, the matcher, the PPM, the fuser, and the backend. Given an input image, a CNN-based encoder neural network extracts multiple feature maps in different scales. The features are fed into the PPM head to extract global features and into the AFPN. The objective of the AFPN is to calculate the window cross-attention between the different feature maps. Then, the AFPN feature maps are passed to the matcher module, which combines the features by concatenating and upscaling the features by transpose convolutions. Thereafter, the features are fed into an adaptive convolutional layer in the fuser module. In order to mix local high-resolution features with global low-resolution features, features from the PPM head are pooled and serve as the kernel weights of the adaptive CNN. Finally, a backend consisting of multiple sequences of CNNs, batch normalization, and activation functions followed by a CNN predicts the dust map

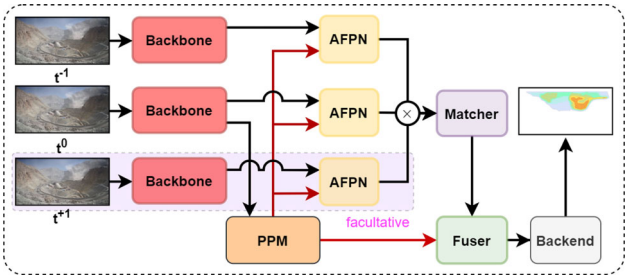

Fig. 13 *DustNet C:* Late multiscale feature fusion

features are processed by the backend, which consists of multiple sequences of CNNs, into a dust map.

Temporal Fusion of DustNet. After investigating the submodules, we focus on the different temporal fusion approaches. DustNet processes input image sequences X of the dimensions $T \times H \times W \times 3$ to a continuous dust density map $\frac{H}{2} \times \frac{W}{2} \times 1$. T may consist of a maximum of three consecutive images, where the target y is assigned to the image x_{t^0}. The images are fed into the backbone, and to leverage the temporal information between consecutive images, DustNet utilizes various approaches.

Concatenation of Images. An obvious way to concatenate the images to $D \times H \times W$ is where the product of the number of images T and the number of channels C is the new channel dimension D. Examples of this approach can be found in Cheng et al. (2021) or Liu et al. (2022b). Hereby the backbones are usually specially adapted to 3D input. This version is called DustNet A.

Early Multiscale Feature Fusion. DustNet B follows a different fusion scheme. Features from three backbones, which share weights, are fed into the temporal merger (TM) neural network. TM subtracts the feature maps of image $x_{t^{-1}}$ and $x_{t^{+1}}$ respectively from x_{t^0} and multiplies the difference. We pass the new feature maps through a 2D pointwise convolutional layer and add skip connections from the feature maps of the image x_{t^0} to the output.

Late Multiscale Feature Fusion. This approach represents a simple fusion of AFPN features (see Fig. 13). Backbone and AFPN weights are shared between the instances. The PPM head is only fed with the backbone features from image x_{t^0}. To lower the computational complexity and potentially minimize convergence issues by reducing the data input, only two consecutive images may be utilized. This version is referred to as DustNet C.

Adaptive Global Information Feature Fusion. The goal of DustNet D is to calculate the global aggregated features from a PPM head for each image. Backbone and PPM weights are shared. The local feature branch is only fed with the multiscale backbone features from image x_{t^0}. The fusion of the temporal information occurs in the fuser module. For each PPM head, an adaptive convolutional layer is added.

Acknowledgements The images in the presented figures and those used for creating the Meteodata dust dataset are from the pit of Minera Los Pelambres, which collaborates with Meteodata in the advanced use of cameras for emission control strategies. The permission to use the images in this publication is kindly appreciated.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The Meteodata dust dataset is not publicly available at the time of publication. In contrast, the URDE dataset is publicly available and can be found under the references (De Silva et al., 2023).

References

Agarwal, A., & Arora, C. (2023). Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5861–5870).

Avvenuti, M., Bongiovanni, M., Ciampi, L., Falchi, F., Gennaro, C., & Messina, N. (2022). A spatio-temporal attentive network for video-based crowd counting. In *Proceedings of the 2022 IEEE symposium on computers and communications* (pp. 1–6). IEEE.

Bhat, S. F., Alhashim, I., & Wonka, P. (2021). Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4009–4018).

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th international conference on pattern recognition* (pp. 3121–3124). IEEE.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv:1604.06174

Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., & Schwing, A. G. (2021). Mask2former for video instance segmentation. arXiv:2112.10764

Cheng, Z.-Q., Dai, Q., Li, H., Song, J., Wu, X., & Hauptmann, A. G. (2022). Rethinking spatial invariance of convolutional networks for object counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19638–19648).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems, 34*, 3965–3977.

De Silva, A., Ranasinghe, R., Sounthararajah, A., Haghighi, H., & Kodikara, J. (2023). A benchmark dataset for binary segmentation and quantification of dust emissions from unsealed roads. *Scientific Data, 10*(1), 14.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transattentions for image recognition at scale. arXiv:2010.11929

Eigen, D., Puhrsch, C., Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, Vol. 27.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks, 107*, 3–11.

FairScale authors. (2021). FairScale: A general purpose modular PyTorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale

Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2002–2011).

gabort@AdobeStock. (2023). https://www.stock.adobe.com

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUS). arXiv:1606.08415

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the international conference on machine learning* (pp. 448–456).

Jain, P., Jain, A., Nrusimha, A., Gholami, A., Abbeel, P., Gonzalez, J., Keutzer, K., & Stoica, I. (2020). Checkmate: Breaking the memory wall with optimal tensor rematerialization. *Proceedings of Machine Learning and Systems, 2*, 497–511.

Khan, A. R., & Khan, A. (2023). Maxvit-unet: Multi-axis attention for medical image segmentation. arXiv:2305.08396

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

Lee, J., Shi, Y. R., Cai, C., Ciren, P., Wang, J., Gangopadhyay, A., & Zhang, Z. (2021). Machine learning based algorithms for global dust aerosol detection from satellite images: Inter-comparisons and evaluation. *Remote Sensing, 13*(3), 456.

Lee, M., Hwang, S., Park, C., & Lee, S. (2022). Edgeconv with attention module for monocular depth estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2858–2867).

Leghari, S. K., Zaidi, M. A., Siddiqui, M. F., Sarangzai, A. M., Sheikh, S.-U.-R., & Arsalan. (2019). Dust exposure risk from stone crushing to workers and locally grown plant species in Quetta, Pakistan. *Environmental Monitoring and Assessment, 191*(12), 1. https://doi.org/10.1007/s10661-019-7825-1

Li, X., Chen, S., Hu, X., & Yang, J. (2019). Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2682–2690).

Li, Y., Zhang, X., & Chen, D. (2018). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1091–1100).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117–2125).

Liu, W., Salzmann, M., Fua, P. (2019). Context-aware crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5099–5108).

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. (2022). Swin transattention attention v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12009–12019).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transattention attention: Hierarchical vision transattention attention using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transattention attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3202–3211).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv:1711.05101

Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H. (2020). Hybrid graph neural networks for crowd counting. In *Proceedings of the AAAI conference on artificial intelligence* Vol. 34 (pp. 11693–11700).

Michel, A., Weinmann, M., Schenkel, F., Gomez, T., Falvey, M., Schmitz, R., Middelmann, W., & Hinz, S. (2023). Terrestrial visual dust density estimation based on deep learning. In: Proceedings of the 2023 IEEE international geoscience and remote sensing symposium.

Michel, A., Weinmann, M., Schenkel, F., Gomez, T., Falvey, M., Schmitz, R., Middelmann, W., & Hinz, S. (2024). DustNet: Attention to dust (pp. 211–226). https://doi.org/10.1007/978-3-031-54605-1_14

Morman, S. A., & Plumlee, G. S. (2014). Dust and human health. Mineral dust: A key player in the Earth system (pp. 385–409).

NASA. (2023). Mars curiosity image gallery. Retrieved March 08, 2019, from https://www.nasa.gov/mission_pages/msl/images/index.html

Patil, V., Sakaridis, C., Liniger, A., & Van Gool, L. (2022). P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1610–1621).

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transattentions for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12179–12188).

Ren, J., Zhang, M., Yu, C., & Liu, Z. (2022). Balanced MSE for imbalanced visual regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7926–7935).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, Vol. 28.

Sam, D. B., Surya, S., & Babu, R. V. (2017). Switching convolutional neural network for crowd counting. In *Proceedings of the 2017 IEEE conference on computer vision and pattern recognition* (pp. 4031–4039). IEEE.

Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. arXiv:1803.02155

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1874–1883).

Song, M., Lim, S., & Kim, W. (2021). Monocular depth estimation using Laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(11), 4381–4393.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Tay, Y., Dehghani, M., Abnar, S., Chung, H. W., Fedus, W., Rao, J., Narang, S., Tran, V. Q., Yogatama, D., & Metzler, D. (2022). Scaling laws vs model architectures: How does inductive bias influence scaling? arXiv:2207.10551

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). Maxvit: Multi-axis vision transformer. In *European conference on computer vision* (pp. 459–479). Springer.

U.S. Environmental Protection Agency. (2022). Particulate Matter (PM) 2022 Report. Retrieved 22 May, 2024, from https://www.epa.gov/system/files/documents/2023-06/PM_2022.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, Vol. 30.

Wang, L., Zhang, J., Wang, Y., Lu, H., & Ruan, X. (2020). Cliffnet for monocular depth estimation with hierarchical embedding loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V* (pp. 316–331). Springer.

Yuan, F., Zhang, L., Xia, X., Huang, Q., & Li, X. (2020). A wave-shaped deep neural network for smoke density estimation. *IEEE Transactions on Image Processing, 29*, 2301–2313.

Yuan, W., Gu, X., Dai, Z., Zhu, S., & Tan, P. (2022). Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3916–3925).

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016) . Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 589–597).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine, 5*(4), 8–36.