Karlsruhe Institute of Technology

# Investigating radar-to-depth cross-sensor gait recognition via GANs

Master's Thesis of

Yannick Engel

At the KIT Department of Informatics
KASTEL – Institute of Information Security and Dependability

| | |
|---|---|
| First examiner: | Prof. Dr. Thorsten Strufe |
| Second examiner: | Prof. Dr. Christian Wressnegger |
| First advisor: | M.Sc. Julian Todt |

23. September 2025 – 23. March 2026

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

*Investigating radar-to-depth cross-sensor gait recognition via GANs (Master's Thesis)*

I declare that I have developed and written the enclosed thesis completely by myself. I have not used any other than the aids that I have mentioned. I have marked all parts of the thesis that I have included from referenced literature, either in their original wording or paraphrasing their contents. I have followed the by-laws to implement scientific integrity at KIT.

**Karlsruhe, 23. March 2026**

Yannick Engel

. . . . . .

(Yannick Engel)

# Abstract

This thesis examines whether GAN-based radar-to-depth translation can enable downstream recognition with an existing depth-based gait-recognition model. This question is privacy-relevant as it would enable gait recognition for radar data without requiring an explicitly identity-labeled radar dataset. To study this setting, a GAN-based translation model is proposed, consisting of a PointNet encoder, a FiLM-conditioned convolutional decoder, and a PatchGAN discriminator. The model is trained on a paired but unlabeled radar-depth dataset and evaluated through downstream recognition with a depth-based gait recognition model. The results show that this radar-to-depth model enables downstream recognition to a limited but measurable extent (Rank-1 accuracy: 5.66%) above the chance level (2.63%). This demonstrates that identity-relevant gait characteristics are preserved during translation at least for some parts. However, it also showcases a substantial gap towards the recognition performance on real depth sequences. Therefore, not a strong recognition attack is established, but a plausible privacy risk pathway is demonstrated that bridges the gap between radar and depth sensors for gait recognition.

# Zusammenfassung

In dieser Arbeit wird untersucht, ob eine GAN-basierte Radar-zu-Tiefen-Translation die nachgelagerte Erkennung mit einem bestehenden tiefenbasierten Gangerkennungsmodell ermöglichen kann. Diese Frage ist datenschutzrelevant, da sie eine Gangerkennung für Radardaten ermöglichen würde, ohne dass ein explizit mit Identitäten beschrifteter Radardatensatz erforderlich wäre. Um diesen Ansatz zu untersuchen, wird ein GAN-basiertes Translationsmodell vorgeschlagen, das aus einem PointNet-Encoder, einem FiLM-konditionierten konvolutionellen Decoder und einem PatchGAN-Diskriminator besteht. Das Modell wird auf einem gepaarten, aber unmarkierten Radar-Tiefen-Datensatz trainiert und durch nachgelagerte Erkennung mit einem tiefenbasierten Gangerkennungsmodell evaluiert. Die Ergebnisse zeigen, dass dieses Radar-zu-Tiefen-Modell eine nachgelagerte Erkennung in begrenztem, aber messbarem Umfang (Rank-1-Genauigkeit: 5, 66%) über dem Zufallsniveau (2, 63%) ermöglicht. Dies zeigt, dass identitätsrelevante Gangmerkmale während der Übersetzung zumindest teilweise erhalten bleiben. Es zeigt jedoch auch eine erhebliche Lücke zur Erkennungsleistung bei echten Tiefensequenzen auf. Daher wird zwar kein starker Erkennungsangriff etabliert, aber ein plausibler Pfad für ein Datenschutzrisiko aufgezeigt, der die Lücke zwischen Radar- und Tiefensensoren für die Gangart-Erkennung überbrückt.

# Contents

# List of Figures

# List of Tables

# 1. Introduction

In our everyday life, biometric data is captured in many ways. This can occur with awareness, through active interaction with a biometric authentication system, or without meaningful awareness, via social media or surveillance. Biometric recognition is widely used for authentication and access control. However, it raises substantial privacy concerns if an instance of a biometric trait can be linked to a specific identity across contexts. This issue is especially relevant for behavioral biometrics. Behavioral traits often exhibit sufficient inter-individual variation to allow identification and can be collected during regular activities such as walking, speaking, or general human-computer interaction [21]. Behavioral biometrics have legitimate use cases but also pose a risk of identity disclosure when reused across systems or in inference settings.

Gait is an example for which these privacy concerns arise. As a behavioral trait, gait describes the characteristic locomotion of humans and contains person-specific information [39]. Previous studies show gait can be captured from a distance without explicit participation of the observed individual [35, 43]. These features make gait recognition attractive for security applications but also raise privacy concerns.

Privacy considerations become more complex as sensing technologies evolve. Sensor selection depends on technical constraints, operational requirements, and data privacy or legal regulations. Camera-based sensors may be restricted or considered undesirable, but other sensors may still be permitted. Radar is a notable candidate. It captures motion without creating conventional visual imagery and continues to work even when cameras are impaired [37]. This creates an apparent privacy advantage. Replacing camera-like sensors with radar may be perceived to reduce the risk of biometric identification.

However, this implication holds only if identity-relevant information cannot be transferred between sensor domains. If a model can convert radar observations into another representation that is usable by a video- or depth-based recognition system, the sensor change may not eliminate the privacy risk after all. Instead, it only changes the technical path for biometric inference. Therefore, the question arises of whether such a cross-sensor recognition is feasible and whether such privacy-friendliness claims can be overcome through data translation.

For mmWave radar and depth cameras, the sensor domains differ vastly in their data formats. Depth cameras produce image-like representations with spatial structure, while mmWave radar produces sparse, noisy point clouds [37]. Previous research on radar-to-depth translation for human pose estimation has shown that radar point clouds can be converted into depth images [14]. However, sparsity and noise in radar point clouds are

challenging, and subtle movements can be lost if point cloud density is too low. On the other hand, cross-sensor gait recognition studies show that depth-like point-cloud projections help reduce the domain gap between video and point-cloud data. However, this applies to dense LiDAR point clouds, not sparse radar point clouds [17].

Existing research has investigated cross-sensor gait recognition across several sensor combinations, and the feasibility of radar-to-depth translation for human pose estimation has been demonstrated. However, it remains unclear whether radar point cloud sequences can be translated into depth-silhouette sequences while preserving identity-relevant gait characteristics. This thesis addresses this gap by investigating radar-to-depth translation for downstream gait recognition. In particular, a privacy-relevant setting is considered in which only paired but unlabeled radar-depth recordings are available, and recognition relies on an existing depth-based gait recognition model.

Related work provides methodological motivation for a generative adversarial approach for this translation. In cross-spectral face recognition, adversarial image generation translates sensor data to the domain used by current recognition systems [2]. In gait recognition, generative adversarial models additionally show that identity-relevant characteristics can be preserved despite changes in viewpoint, appearance, or framerate [50, 52].

The contributions of this thesis can be broken down into two parts. First, a GAN-based radar-to-depth translation pipeline for downstream depth-based gait recognition is proposed. Second, an extensive evaluation of the proposed method, including a comprehensive ablation study and investigation of result variability, is performed.

To achieve this, the thesis is structured as follows. Section 2 introduces gait as a behavioral biometric, gait recognition, radar sensing, and generative adversarial networks as the foundations of this thesis. In Section 3, related work is discussed, and the thesis is positioned with regard to the identified research gap. In Section 4, the proposed approach is explained, starting from the abstract adversary model and proceeding down to architectural design choices. Section 5 then systematically evaluates the proposed model with respect to the concrete research question but also with respect to specific model behavior. Afterwards, in Section 6, the main findings are discussed, including the corresponding implications and existing limitations of this thesis. Finally, Section 7 concludes the thesis.

# 2. Background

This chapter introduces the methods and concepts that are used as a foundation for this thesis. Section 2.1 introduces biometrics and gait as a behavioral biometric. Section 2.2 describes mmWave radar as a sensing technology and introduces PointNet as a model to process point cloud data. Section 2.3 covers gait recognition for different sensor data types. Afterwards, Section 2.4 introduces the concept of cross-sensor recognition and describes different categories of approaches for this problem class. Section 2.5 explains generative adversarial networks and the extensions on that defined architecture. Finally, Section 2.6 introduces overlap-based metrics used in the loss function and evaluation.

## 2.1. Biometrics

Biometrics are defined as physiological or behavioral characteristics that can be used to identify a person or verify claimed identity. Biometric recognition typically refers to the automated process of recognizing individuals using their biological characteristics, with two core tasks: verification and identification [21]. Verification is the process by which a person claims an identity, and the biometric attributes are checked against it. Identification, on the other hand, concerns selecting the concrete identity of a person from a set of known identities. While this describes the closed-set definition, the open-set definition also allows the result to be unknown if the required confidence is not achieved.

Biometrics are divided into two categories. The first is physiological traits. This includes well-known biometrics, such as fingerprints, faces, and irises, which are widely used for recognition tasks. The other class is behavioral traits. These include, but are not limited to, keystroke cadence, mouse movements, voice patterns, and gait patterns. While they can change over time and be actively influenced, they also produce individual, consistent patterns subconsciously. Whereas physiological traits can be captured with momentary data acquisition alone, behavioral traits require observation over time to even extract the temporal dependencies needed for identifying patterns. However, regardless of the specific category of biometric information, all of it can enable unique identification and can be used to derive other sensitive information about an individual [21].

### 2.1.1. Gait

Gait is a behavioral biometric and refers to the characteristic pattern of human locomotion, varying with the type of movement, e.g., walking or running. It is considered a largely

automated motor task, governed by coordinated neuromuscular control [39]. A gait cycle (also called a stride) is one complete sequence of movement, beginning with the initial ground contact of a foot and ending when the same foot contacts the ground again. Stöckel et al. [39] indicate that this pattern has unique features for each individual, leading to the potential of gait characteristics being used for recognition purposes.

Compared with other biometric features used for recognition, such as fingerprints, faces, or mouse movements, gait has special characteristics: It can be captured from a distance without requiring the cooperation of the observed person. Additionally, simple instruments can be sufficient, and low resolution is less of a problem than with other biometrics. Furthermore, it is difficult to conceal one's own gait pattern and to impersonate another person's gait pattern. All of these make gait a well-studied candidate for recognition tasks [43].

## 2.2. Radar

Millimeter-wave (mmWave) radar is a sensing technology that operates at high-frequency radio waves (millimeter wavelengths) to detect objects and estimate their locations and motion [37].

The core idea is that the radar continuously transmits a signal whose frequency increases linearly over time (a chirp) and then listens to the reflected signal, which arrives slightly delayed. By mixing the original chirp with the received signal to an intermediate frequency, range, radial velocity, and angle can be estimated. This, therefore, allows mapping those detections to a spatial coordinate system, giving an approximate 3-dimensional location [37].

Two commonly used formats for mmWave radar data are micro-Doppler spectrograms and radar point clouds. Micro-Doppler spectrograms are time-variant Doppler signatures that capture micro-motions. Radar point clouds, on the other hand, are lists of detected points after all signal-processing steps, which include 3D coordinates, velocities, potential intensities, and reflection values. While micro-doppler spectrograms capture motion patterns over time, radar point clouds directly capture geometric characteristics [37].

Although mmWave radar is a good solution for detection, it also has its issues, mainly due to noise. One big example is multipathing. This describes the phenomenon in which a signal is returned not only directly, but also indirectly via reflection from another object. This results in a detected ghost point where no detection should occur. Additionally, environmental interference and potential micro-movements within the environment can lead to noisy detection [37].

However, mmWave radar also offers advantages compared to other sensor technologies. It is suitable in low-light environments and is robust to adverse weather conditions, both of which are issues for video-based sensing. Additionally, it can capture a person's motion and geometry without recording conventional visual recordings of identifiable facial features [37]. This has therefore been discussed as an alternative to privacy-invasive video-based sensing.

### 2.2.1. PointNet

Working with point cloud data, especially in the machine learning field, produces its own set of challenges. One being directly influenced by the irregularity of this data format. Those stand out when comparing point clouds to images. While images, at least within the same dataset, have the same resolution and the data for each image is ordered, this is not necessarily the case for point cloud data of any kind. Point cloud data typically varies in size depending on the scenario and the number of detected points, and is not necessarily ordered. Nevertheless, point cloud data is used for machine learning purposes, introducing the need for an encoder that works with varying input sizes but produces one fixed latent representation, invariant to permutation of the points.

Qi et al. [28] aimed to address this with PointNet. PointNet is a deep learning architecture that directly consumes point clouds as unordered point sets, producing a suitable encoding for classification and part/scene semantic segmentation. Thus, PointNet provides an option for encoding point cloud data that avoids converting it into voxels or multi-view images, which can be bulky and inefficient.

The design of PointNet is theoretically grounded to address issues related to varying point sizes and permutation invariance. The core insight Qi et al. [28] propose, is that any continuous function over a point set can be approximated by applying a shared transformation $h$ independently to each point, followed by a symmetric function $g$ that aggregates the results:

$$f(\{x_1, \ldots, x_n\}) \approx g(h(x_1), \ldots, h(x_n))$$

In practice, $h$ is implemented using a shared multi-layer perceptron (MLP), which is applied independently to each point. The global aggregation $g$, on the other hand, is achieved by a global max-pooling operation. Due to the order-invariance of both operations, the resulting function $f$ is independent of ordering and can handle varying input sizes. Additionally, the approximation guarantee ensures that the architecture is expressive enough to approximate any continuous set function on a point set. The overall architecture is shown in Figure 2.1.

To achieve input invariance to geometric projections such as rotations, PointNet introduces two transformation networks. The first transformation network operates on the raw input point cloud and predicts a transformation matrix that is applied to the point cloud. Its goal is to align the input point cloud with a canonical representation to improve robustness to geometric transformations. The second transformation network takes the intermediate feature vector after the first MLP stage and applies an equivalent alignment in feature space. Both transformation networks mirror the overall PointNet architecture in itself and are identity-biased during training for stability Qi et al. [28].

PointNet was introduced specifically for classification and segmentation. For classification, the global feature vector is used directly as input to the classification network. However, for segmentation tasks, the 64-dimensional intermediate feature of each point is used, and the global feature is appended to each point, since the segmentation task produces an output for each point separately Qi et al. [28].

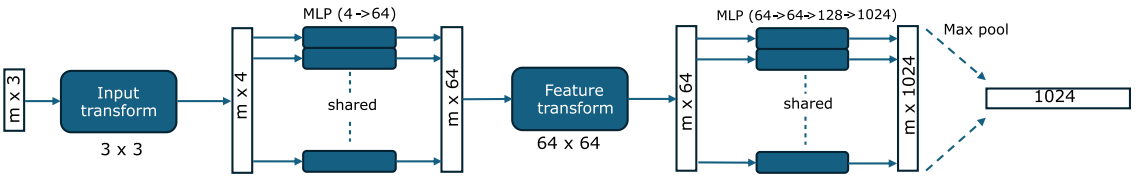



**Figure 2.1.:** Model Architecture of PointNet

While their paper mainly examined segmentation and classification, PointNet has since been applied to many more applications. Cheng and Liu [6] showed the use of PointNet for radar point clouds in the context of gait recognition, in which they used global features as an encoding of a single frame. Milz et al. [24], on the other hand, use PointNet global features of a LiDAR point cloud in addition to a 2D-mapped version of the same point cloud, among other things, to generate realistic-looking images of the scene at hand. Therefore, demonstrating the successful application of PointNet beyond strict object classification and segmentation tasks.

One known limitation of PoinNet is that it does not capture local geometric structures well because each point is processed individually. PointNet++ [29] addresses this issue by applying PointNet recursively over nested local neighborhoods.

## 2.3. Gait Recognition

Gait recognition aims to derive the individual's identity from their gait [43, 32]. In gait recognition, the goals can be divided into three categories. The first is identification. Here, the goal is to select the closest matching identity from a closed set of known identities for the observed gait. Another task is verification. Here, the decision is limited to two outcomes: accepting the observed gait or rejecting it. Finally, there is re-identification (ReID). Here, the goal is to retrieve the same subject across different cameras, times, or even sensors, connecting those observations to a single identity without requiring knowledge of the observed identity.

As mentioned, gait recognition is not limited to a single sensor type. Different sensors have been shown to perform gait recognition successfully. Most notably, different kinds of video data and LiDAR data produce a major part of the scientific research regarding gait recognition [43, 32]. However, other sensors such as radar [54] or WiFi [5] have also been shown to be sufficient for gait recognition. While all of these are appropriate for gait recognition, they can differ widely in their approaches depending on the data.

For video data, the main focus is on the persons themselves and on removing the possible influence of background or texture. The approaches can be generally divided into two paradigms: model-based and model-free approaches [43]. Model-based approaches aim to explicitly model the human body structure to extract concrete features. This can range from

simple calculations of specific gait features, such as stride length and cadence of the input video sequence, to modelling the body explicitly by extracting key points that are tracked over time and used to derive gait information. These approaches yield procedures that are independent of scale and viewpoint. However, they strongly rely on the quality of the video input [43]. Model-free approaches, on the other hand, either consume the silhouette sequence directly or use aggregated silhouette templates to implicitly learn relevant gait features. The use of silhouettes is a standard approach because it reduces gait-unspecific influences. If not used as a sequence, the silhouettes are aggregated into a so-called gait template to compress the sequence information into a single representation. Compared with model-based approaches, model-free approaches are less dependent on good video quality. However, they are, as expected, much more reliant on view and scale [43].

LiDAR data is available as 3D point clouds. A recording of a person's gait pattern is therefore a sequence of point clouds. For LiDAR used as input to gait recognition, the point cloud data is typically first mapped to a lower-dimensional intermediate representation such as depth maps [33]. As point clouds are 3D data, they enable mapping them into depth representations from different perspectives. This enables elevating the approach from single-view to multi-view [33]. Thereby, enriching the available information and creating a format that is structurally consistent and more CNN-friendly.

Gait recognition using radar data can be divided into approaches based on the data format. When using micro-Doppler spectrograms, the approaches are similar to video-based approaches, in which each micro-Doppler signature corresponds to a single frame. Therefore, in a model-free approach, these frames would be consumed directly [8], while a model-based approach would try to extract explicit features such as stride rate [43]. However, when using radar point clouds, gait recognition must be handled differently, as point clouds do not provide a consistent data format in terms of size. An existing approach is therefore to first encode each radar point cloud frame with a point cloud-specific encoder, such as PointNet. Afterwards, the consecutive frame-wise encodings are aggregated to capture the spatio-temporal features present in the gait sequence [6]. However, the specifics of radar point clouds have to be respected. Radar point clouds not only consist of 3-dimensional coordinates but also include additional per-point information, such as velocity and reflectance. When encoding point clouds, this information must be handled separately. Operations such as transformations in the encoder may be appropriate for coordinates but not for other attributes of each point [6]. Additionally, the sparsity and noisiness of radar point cloud data has to be respected [37],

## 2.4. Cross-sensor Recognition

Cross-sensor recognition can be described as a recognition task in which data is collected from different sensors. Here, sensor refers broadly to a data acquisition modality or device. Even if both sensors nominally measure the same phenomenon, they may induce a domain gap, even if it is just a difference in viewpoints. This leads to an issue that models trained on one sensor may fail to generalize to another without explicit alignment or adaptation.

However, the nomenclature of "cross-sensor" is not standardized and can also be found as "cross-modal" [7, 44, 36] or more adapted to a concrete sensor domains, e.g., cross-spectral if both sensors work with different spectra of light [2].

One way to differentiate existing approaches is to consider the availability of sensors at inference: single-sensor or multi-sensor inference. Single-sensor inference refers to approaches that, as the name suggests, operate when only one sensor is available at inference. Here, mainly two different types exist. The first is a representation-alignment approach, in which the goal is to learn an invariant embedding on which recognition can be trained or used for re-identification [7, 44]. Therefore, regardless of the available sensor, an embedding can be created and used for recognition. The second is a translation-based approach in which one sensor domain is translated into the other, enabling recognition in the target domain [51, 4]. For multi-sensor inference, the approach depends on the availability of each sensor at inference time. The architecture fuses the different inputs at any stage to potentially complement one another and compensate for the weaknesses of a single sensor.

## 2.5. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [15] are generative models that learn to generate samples that match the distribution of an existing dataset. Compared with discriminative models that learn a concrete mapping from input to output, GANs aim to learn the underlying data distribution $p_{data}$ of the original dataset. While generative models are not new, they can struggle to approximate complex probabilistic computations arising in maximum likelihood estimation. GANs aim to address this issue by introducing an adversarial training mechanism to eliminate the need for explicit likelihoods.

A GAN consists of two separate networks. A Generator ($G$) and a Discriminator ($D$). The Generator takes a noise $z$ as input and maps it to the target space to create a synthetic sample $G(z)$. The Discriminator, on the other hand, gets a sample $x$ as input and outputs a scalar according to the guessed probability whether the sample $x$ is real and therefore from the original data distribution $p_{data}$ or not. It provides a training signal by evaluating the Generator's output samples. While the networks have different goals, they mutually improve by competing. If $D$ improves, it can generate informative feedback for $G$, and if $G$ improves, $D$ must fine-tune its decision boundary, thereby contributing to each other during training [15].

This interaction is formalized as a minmax objective:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

where $p_{data}$ is the real data distribution, $p_z$ is the distribution of the noise input of $G$. $D(x)$ and $D(G(z))$ correspond to the output of D depending on the respective input [15].

On the one hand, it is the goal to maximize the output of D on real data, while on the other hand, one minus the output on the generated data should be minimized Both of these goals

are contrary to each other and would lead to a theoretical limit of D only being able to guess if $p_{data} \equiv G(z)_{z \sim p_z}$ [15].

### 2.5.1. Conditional Generative Adversarial Networks

While GANs aim to learn to generate new samples from an existing data distribution, this alone is not sufficient for some tasks. The sample that should be generated may depend on additional information, such as an attribute or characteristic. Conditional Generative Adversarial Networks (cGAN) [25] aim to explicitly introduce this option of control. For cGANs, a condition variable $c$ is introduced that is consumed by both the Generator and the Discriminator. With this, $G$ aims not to learn the data distribution $p_{data}(x)$ but the data distribution conditioned on one class or attribute $p_{data}(x|c)$. This changes the minmax objective to the following

$$\min_G \max_D V(G, D, c) = \mathbb{E}_{x \sim p_{data}}[\log D(x|c)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z|c)|c))]$$

The condition is required in the Generator to generate a sample that satisfies the condition, and the Discriminator needs the condition to determine which conditioned data distribution the sample should come from [25].

With the extension of the input of the Discriminator with the condition, it introduces multiple possibilities to embed the condition into the architecture of the Discriminator. Two straightforward approaches are to concatenate the condition to the discriminator's input or to any hidden layer in the discriminator. A third option is introduced by Miyato and Koyama [26]. They propose a projection-based conditional embedding in the form of:

$$D(x, c) := \psi(\phi(x)) + c^T V \phi(x) \tag{2.1}$$

where $\phi$ describes the discriminator without the last layer, $\psi$ represents the final output layer, and $V$ denotes a projection matrix that is trained jointly with the rest of the network. The output of the Discriminator then consists of the inner product of the projected condition with the features before the last layer, added to the original unconditioned discriminator output. This adaptation has been shown to empirically outperform previous conditioning approaches for conditional image generation and super-resolution [26].

One paradigm that evolved from cGANs is referred to as Mapping-GANs. There, the condition is not a label but a concrete sample from the source distribution, with the goal of learning a mapping between the two domains. This paradigm is exemplified in image-to-image translation [20]. In this setting, an input image serves as the conditioning signal, and the generator learns to produce a corresponding output image in the target domain. The task of the discriminator is then to decide not only whether the output is realistic but also whether it fits the input image condition. Additionally, the noise vector can be completely omitted, as is often the case when deterministic paired training is used.

A limitation of a global discriminator for an image generation task is that a single score for the entire image produces only a weak training signal for local structures. PatchGAN [20] approaches this issue by outputting a spatial map of local scores. These aim to classify whether a corresponding $N \times N$ patch of the input image is real or generated, rather than providing a single value for the entire image. This leads to improved sharpness of local structures, thereby generating a better feedback signal for image generation tasks.

### 2.5.2. Least-Squares Generative Adversarial Networks

Standard GANs typically use sigmoid or binary cross-entropy for the discriminator loss. However, this can lead to vanishing gradients for G when D becomes too confident. Least-squares generative adversarial networks (LSGAN) [23] aim to mitigate this issue. The logistic cross-entropy loss is replaced with a least-squares loss, which accounts for the distance to the concrete decision boundary. It improves the overall quality of the Generator by contributing a feedback signal even if the decision is correct. This creates two new objective functions to optimize in the conditional context:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{data}}[(D(x|c) - \beta)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z}[(D(G(z|c)|c) - \alpha)^2]$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z\sim p_z}[(D(G(z|c)|c) - \gamma)^2]$$

where $\alpha$ is the label for fake data, $\beta$ is the label for real data, and $\gamma$ describes the value that G wants D to produce of the generated input (often $\alpha = 0$, $\beta = 1$, and $\gamma = 1$). This leads to a discriminator loss that can produce a significant gradient even for samples far from the decision boundary, thereby mitigating potential issues with vanishing gradients.

### 2.5.3. Feature-wise linear modulation

Many generative architectures require a mechanism for injecting a signal into the processing of another. Simple approaches, such as concatenation at the bottleneck stage, can exert only a limited influence throughout the network. Feature-wise linear modulation (FiLM) aims to address this by providing a general-purpose conditioning layer that can be inserted at arbitrary network layers [27].

FiLM learns a projection of the condition onto per-channel scale and shift parameters $\gamma$ and $\beta$. These are then applied via a feature-wise affine transformation:

$$FiLM(F_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}F_{i,c} + \beta_{i,c} \tag{2.2}$$

The conditioning input can be any differentiable signal, ranging from language to embedding or sensor data. This makes FiLM applicable across a wide range of tasks. While it was originally introduced for visual reasoning [27], it has since been applied to other fields, such as text-conditioned image manipulation [16].

## 2.6. Metrics

This section briefly introduces the overlap-based metrics used in the loss function and in the evaluation of this thesis. Both metrics can measure the spatial overlap between a generated and a ground-truth mask.

### 2.6.1. Dice coefficient

The Dice coefficient is a similarity measure used to quantify overlap in segmentation scenarios. It was introduced independently by Dice [11] and Sørensen [38] to quantify overlap between two sets. It is defined as:

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

where $X$ and $Y$ describe two different sets and $Dice \in [0, 1]$. This metric has become widely used in image segmentation to measure the quality of a predicted mask [12].

For segmentation tasks, where the mask is continuous rather than binary, Dice can simply be calculated with the following equation:

$$Dice(x, y) = \frac{2 \cdot \sum_i x_i \cdot y_i}{\sum_i x_i + \sum_i y_i}$$

where $x_i$ and $y_i$ represent the $i$-th entry of the ground truth mask $x$ and the predicted mask $y$, respectively. This definition avoids a non-differentiable threshold step and preserves gradient flow during training.

### 2.6.2. Tversky Index

Tversky [42] introduces a similarity measure, in which common and distinctive features can be weighted asymmetrically. The idea is relevant in segmentation settings, where false positives and false negatives may carry different levels of importance. Therefore, it generalizes overlap-based similarity measures by allowing separate weighting of false positives and false negatives [12]. This represents a more generalizable version of the Dice coefficient. The Tversky index is defined over two sets $X$ and $Y$ as:

$$Tversky_{\alpha,\beta}(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|Y \setminus X| + \beta|X \setminus Y|}$$

where $\alpha$ and $\beta$ describe the weighting of false positives and false negatives respectively. This allows an asymmetric approach to the weighting of false positives and false negatives. The special case $\alpha = \beta = \frac{1}{2}$ reduces Tversky to the Dice coefficient.

As with Dice, Tversky can be used for segmentation on continuous values and a continuous differentiable form can be defined:

$$Tversky(x, y)_{\alpha,\beta} = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i \cdot y_i + \alpha \cdot \sum_i (1 - x_i) \cdot y_i + \beta \sum_i x_i \cdot (1 - y_i)}$$

where $x_i$ and $y_i$ represent the $i$-th entry of the ground truth mask x and the predicted mask y respectively.

# 3. Related Work

Gait Recognition, as a biometric task, has already been widely studied. Current surveys systematize basic approaches, deep models, datasets, and key challenges related to viewpoint/clothing variations, occlusion, and recording conditions [34, 43]. While vision-based approaches were initially strongly characterized by silhouette sequences or gait templates, recent research investigates the feasibility of gait-based identification for non-visual sensors such as LiDAR, Radar, or WiFi. In this thesis, the concrete performance of a single sensor domain is less significant than the question of how gait-relevant information can be transferred or synthesized across sensor domains to enable existing downstream tasks. In this chapter, we review relevant work on cross-sensor gait recognition for different sensor combinations, approaches that combine depth and radar sensors, and related work on generative translation tasks.

## 3.1. Cross-Sensor Gait Recognition

Many relevant cross-sensor gait recognition studies address the modality gap between video-based representations and LiDAR-based point clouds [46, 44, 17, 9]. There, the approach that links domain-specific encoders to produce a fused identity or feature space dominates. The aligned feature space can then be used directly for recognition tasks.

Wang et al. [44] explicitly state the challenge that camera and LiDAR data not only have different resolution and structure, but they also capture different kinds of information. CrossGait [44] addresses this issue by using modality-shared attention and feature adaptation to learn a consistent representation that can be used for cross-sensor matching. Wang et al. [46] also adopt a bridging approach by projecting LiDAR point clouds onto a depth-map-like representation. Afterwards, features are extracted from the LiDAR projection and video-derived silhouettes separately and aggregated into a modality-independent identity embedding. Guo et al. [17] presents another approach that uses camera-based silhouettes and LiDAR point clouds as input and learns to align the feature spaces of the two modalities.

All of the above cross-sensor gait recognition approaches demonstrate that this fusion approach is applicable to point cloud and video data. However, all of these approaches require a labeled dataset to construct their supervised classification and positive-negative relations across samples. This may be hard to acquire, depending on the scenario. Their focus is on supervised cross-sensor recognition, in which cross-sensor identity relations are explicitly learned during training and mapped to a single representation. Therefore, they show that cross-sensor gait recognition is possible, but they do not address whether

one sample from one sensor domain can be translated into the input sensor domain of an already existing recognizer. This gap remains unanswered by those works.

Cross-sensor approaches also exist for radar in combination with video sensors. Shi et al. [35] and Deng et al. [8] demonstrate two different approaches, although both use micro-Doppler spectrograms as their radar input format. While Shi et al. [35] take gait energy images as the complementary input, Deng et al. [8] use optical flow. However, both improve upon radar-only or video-only recognition, indicating that they contain complementary gait information. However, they aim to achieve exactly that: higher accuracy than single-sensor recognition, provided that both input domains are available. Therefore, they do not address our research question, which aims to enable recognition in one domain using translation.

## 3.2. Radar and Depth

For radar and depth data combinations, only a comparatively limited body of research on gait recognition exists. One relevant contribution is that of Cao et al. [3], which investigate cross-sensor gait re-identification across non-overlapping sensor fields using RGB-D cameras and radar sensors. Methodologically, they perform a physically motivated radar point-cloud simulation by constructing a 3D mesh of the human body and synthesizing candidate radar points. These are then cross-referenced with embeddings of real radar point clouds to successfully match identities across different sensors. While they investigate sensor-domain translation, they translate from an information-dense representation to an information-sparse representation. This allows, in their case, for an explicit calculation of potential radar points rather than training a model to achieve this goal. However, they do not aim to translate one domain into the other to enable gait recognition, as their approach is similar to embedding alignment, with the additional step of first computing synthetic radar points. This allows assigning different observations to the same identity, but it does not use translation for downstream recognition.

Fan, Rao, and Wang [14] introduce a radar-to-depth approach focusing on human bodies. Their goal is to develop a multi-modal co-learning setup for human pose estimation that uses radar to generate synthetic depth images. They motivate this approach by using synthetic depth images as auxiliary input in environments where depth images are unreliable. This paper demonstrates that radar point clouds are sufficient to produce corresponding depth images of human bodies. However, similar to the work of Cao et al. [3], they split their approach into two parts. First, they train a translation model, and then train the desired model on these inputs. Therefore, they move the core interest not to the translation but to the downstream model. This is exactly the opposite of our research question. Additionally, a translation that is suitable for human pose estimation is not necessarily suitable for gait recognition, as human pose estimation is relatively static and gait recognition depends on temporal context.

## 3.3. Generative Translation

A relevant paradigm for cross-sensor translation is conditional image generation. Isola et al. [20] propose a general-purpose framework for learning mappings between image domains using cGANs. A key finding is that the combination of adversarial and reconstruction loss produces outputs that are locally sharp and globally consistent. This property is relevant to any translation task in which both fine-grained structure and overall geometry matter. However, whether this paradigm extends to radar-to-depth translation and whether the resulting synthetic depth map preserves structure relevant to downstream gait recognition are not addressed by the image-to-image translation literature.

Within gait recognition, generative adversarial networks provide additional motivation for this thesis. GANs have been shown to serve as models for handling variations in viewpoint, clothing, or frame rate. GaitGAN [52] is proposed to transform observed gait from varying views into a canonical side-view representation. Frame-Gan [50], on the other hand, addresses low-frame-rate gait recognition by generating additional frames depending on the previous ground truth frame. While these works do not address cross-sensor recognition from radar to depth, they show that GANs can be applied to gait-specific tasks and retain gait-discriminative information in the process.

Another methodologically related instance exists in cross-spectral face recognition. Recent surveys distinguish the approaches into two main categories: feature-based matching and cross-spectrum image generation [2]. In the latter, images from one sensor domain, e.g., infrared, are translated into representations in the visible domain, enabling face recognition systems trained on the visible spectrum to be successfully used. This does not solve the research question because radar-to-depth translation has different structural constraints, and face recognition cannot be directly compared to gait recognition. Additionally, these approaches often require a labeled dataset or auxiliary models to condition the loss on retention of attributes [19, 10], identities [53], or facial landmarks [47] in the generated images. This, in itself, is a strong assumption. Nevertheless, those provide a strong methodological analogy. When the inference modality is incompatible with the existing recognizer, a GAN-based translation into the recognizer's input domain is an alternative to training a new recognizer.

As discussed in Section 3.2, Fan, Rao, and Wang [14] establishes the feasibility of the translation but within a pose estimation framing that leaves the gait recognition question open. The question of whether such translations can be learned from paired but unlabeled data that are sufficient for a downstream gait recognition task remains unaddressed in the literature.

Considering this, the literature suggests that GANs are a viable approach to addressing the problem addressed in this thesis. Conditional image translation establishes this approach as a general paradigm, gait-adapted GANs show that gait-related information can be retrained during generation, and cross-spectral approaches demonstrate that recognition via translation is a meaningful strategy in a biometric setting. However, the application to the concrete problem in the thesis cannot be assumed and remains unanswered.

# 4. Methodology

This chapter describes the design choices for a GAN-based approach to gait-identity-preserving radar-to-depth translation, starting from high-level adversary assumptions and progressing to technical realization. Section 4.1 defines the adversarial model that motivates the research question. Afterwards, Section 4.2 describes the approach in which the adversary uses its capabilities and access to achieve its goal and how it relates to alternative approaches. Section 4.3 formulates a formal description of the Problem that needs to be solved. Section 4.5 describes the preprocessing steps applied to the raw radar point cloud and depth recordings to construct a suitable form for training. Section 4.6 describes the model architecture, covering the PointNet-based encoder, the FiLM-conditioned decoder, and the projection-based conditional discriminator. Lastly, Section 4.7 defines the training objectives, which combine reconstruction, adversarial, overlap-based, and gait-embedding losses.

## 4.1. Adversary Model

The adversary's goal is to identify individuals inside camera-restricted areas by linking radar observations to known identities. Since gait is a behavioral biometric that can be captured at a distance and without active subject cooperation, this extends surveillance capabilities beyond the originally intended sensing scope. It enables cross-sensor identity linkage and potentially movement tracking despite the absence of permitted camera-based sensing.

The adversary is assumed to be able to collect radar point cloud recordings of an individual using an mmWave radar sensor at inference time. Additionally, the adversary can query a pre-trained depth-based gait recognition model using depth silhouette inputs and receive an output in the model's embedding space.

In terms of access, the adversary has access to a paired but unlabeled dataset of radar and depth recordings with no identity labels or supervision beyond the recording-level pairing. The adversary also has access to a pre-trained depth-based gait-recognition model trained on depth-silhouette recordings and to depth-silhouette recordings of target individuals.

This set of assumptions may apply to a facility that already operates a depth-based entrance-control or identification system. While radar sensors are permitted in internal areas, cameras may not be allowed for regulatory or privacy reasons. The adversary, therefore, does not need to build a fully new radar-based recognition system. In such a scenario, a paired

radar-depth dataset could arise naturally from sensor testing or controlled comparison recordings, while the depth gallery is acquired from the pre-existing depth recognition structure.

## 4.2. Approach

The proposed attack is based on the idea of adapting the radar data to the recognizer. Concretely, a generative mapping from radar point-cloud recordings to synthetic depth-silhouette sequences is learned. These generated depth sequences are then processed by the existing depth-based gait recognition model without modifying its architecture or retraining.

This generative translation design is chosen over other solution families described in Section 3. A first alternative for identifying people using radar data is radar-native gait recognition. Such an approach would require training a new dedicated recognizer directly on the radar data [54, 6]. However, this would shift the problem from cross-sensor recognition to radar-based recognition. While radar-only gait recognition is feasible, it does not leverage the attacker's access to an existing depth-based recognition model and requires a labeled radar dataset. This, therefore, does not align with the adversary's access, as it imposes stronger assumptions.

The second alternative is direct cross-sensor recognition through a shared embedding space. This is a dominant strategy in camera-LiDAR gait recognition by employing two-stream architectures, feature adapters, and contrastive or metric-learning objectives. Those approaches aim to align heterogeneous representations across modalities in feature space [46, 44, 17]. While they work with point clouds and camera-based inputs, they explicitly require a labeled dataset. These are needed because their objective relies on identity labels to construct supervised classification and positive-negative relations across samples and modalities during training. Additionally, these approaches are defined for LiDAR point clouds. Since LiDAR and radar data differ substantially in spatial density and noise characteristics, approaches developed for LiDAR-camera data cannot be directly applied to radar point clouds without adaptation and validation. Therefore, these approaches cannot be applied directly. Nevertheless, one example of this alternative exists for a radar RGB-D combination. While they do not explicitly align the inputs in a shared embedding space, both inputs are mapped to embeddings in the radar point-cloud feature space [3]. However, as before, stronger assumptions are made about the adversary's access. First, not only depth data but also RGB-D data is required for their 3D-mesh generation algorithm. Second, they require an identity-labeled dataset because they use triplet loss with positive and negative examples during training.

The third alternative recognition architecture is multi-sensor fusion. Radar-vision fusion methods demonstrate that radar and visual modalities provide complementary gait information and can improve recognition when both are available at inference [8, 35]. However, this does not fit our adversary model. First, they have access to micro-Doppler spectrograms

instead of point cloud data. Second, both inputs must be accessible at inference time, as they are required to create the corresponding embedding.

The most relevant methodological related work is therefore cross-sensor translation. In radar-depth sensing, this has been shown for human pose estimation, where radar point clouds are converted into depth images [14]. However, as before, stronger assumptions are made about data access, as a labeled dataset is required to train the complete model in which the translation component resides. Therefore, this approach is also unsuitable, as our adversary's access does not satisfy these assumptions but still demonstrates the feasibility of the underlying translation problem.

The proposed method, therefore, follows a translation-first strategy. During training, a generative model is learned from paired radar and depth recordings. During inference, only radar is required. The generator produces a synthetic depth sequence, which is passed to the pre-trained depth-based gait recognition model. The resulting embedding is then compared against a gallery constructed from real depth recordings of identities of interest. In this way, the modality gap is handled before recognition rather than within the recognizer itself, thereby removing the need for an identity-labeled dataset.

Additionally, a frame-by-frame approach is chosen, mapping radar point-cloud frames to individual synthetic depth frames. This design choice is justified by the alternative being a sequence-level generation approach, such as vid2vid [45]. There, the conditioning for the output frame is extended by the previously generated frame via optical flow warping. This, however, introduces a dependency chain. Errors in early generated frames propagate forward, making the quality of every subsequent frame conditional on the previous ones. This approach does not resolve the fundamental challenge of generating structurally faithful depth images. Therefore, the less complex frame-by-frame generation variant is chosen.

Furthermore, because an existing gait recognition model is used, we can concretize our approach. First, we do not have to generate the complete depth image, including the background. This is not required for the gait recognition task, as it focuses only on the silhouettes present in the frame. Secondly, the gait recognition operates on the sequence of centered and cropped silhouettes. This also reduces complexity by lowering the frame resolution, removing the imbalance between the silhouette and the background, and eliminating the need to correctly position the silhouette in the frame for the generation task.

Conceptually, this strategy is close to cross-spectrum image synthesis for face recognition in which a non-visible-spectrum image is translated into a visible-spectrum representation that can then be processed by an existing recognizer [2]. The present thesis adopts the same high-level logic, but for radar-to-depth gait translation under different sensing and structural constraints.

## 4.3. Problem Definition

Building on the adversarial setting described above, let $\mathcal{P}_F$ denote the space of radar point-cloud frames and $\mathcal{D}_F$ the space of depth-silhouette frames. Furthermore, let $\mathcal{P}$ and $\mathcal{D}$ denote the corresponding spaces of radar and depth-silhouette sequences.

The adversary is given a paired dataset of synchronized radar and depth recordings $\{(R_i, D_i)\}_{i=1}^{N}$, where each radar recording consists of a sequence of radar point-cloud frames and each depth recording consists of the corresponding synchronized depth-silhouette frames.

$$R_i = (r_{i,1}, \ldots, r_{i,T_i}) \in \mathcal{P}$$

$$D_i = (d_{i,1}, \ldots, d_{i,T_i}) \in \mathcal{D}$$

The pairing is established at the recording level through synchronized capture and induces frame-level correspondence between $r_{i,t}$ and $d_{i,t}$.

The generator is therefore defined as a frame-wise mapping

$$f : \mathcal{P}_F \to \mathcal{D}_F,$$

such that, for an input radar frame $r_{i,t}$, the generated output $\hat{d}_{i,t} = f(r_{i,t})$ approximates the corresponding target depth-silhouette frame $d_{i,t}$ while preserving the structure that is relevant for gait recognition.

For a complete radar sequence $R_i$, the frame-wise generator creates a synthetic depth-silhouette sequence

$$\hat{D}_i = (\hat{d}_{i,1}, \ldots, \hat{d}_{i,T_i}), \qquad \hat{d}_{i,t} = f(r_{i,t}).$$

In addition, a pre-trained depth-based gait recognition model

$$\mathcal{R} : \mathcal{D} \to \mathcal{E}$$

is available, where $\mathcal{E}$ denotes the embedding space of the recognizer. The objective is that the reconstructed sequence $\hat{D}_i$ preserves sufficient gait-discriminative structure for the downstream recognizer. More specifically, the embedding $\mathcal{R}(\hat{D}_i)$ should be comparable to the embedding of a real depth-silhouette sequence of the same identity.

Therefore, success is defined by downstream recognition accuracy rather than by frame realism. A generated sequence is considered useful only if it enables correct retrieval of the underlying identity when matched against a gallery of known depth-silhouette recordings in the recognizer's embedding space.

## 4.4. Model Pipeline

This section describes the complete pipeline, from sensor data to downstream recognition. This aims to describe the interactions among all components before describing them in the following sections in more detail. The pipeline comprises sensor-modality-specific preprocessing, training of the generative translation model, inference from radar point-cloud data, and identification in the embedding space of the downstream recognition model. The complete pipeline is visualized in Figure 4.1.

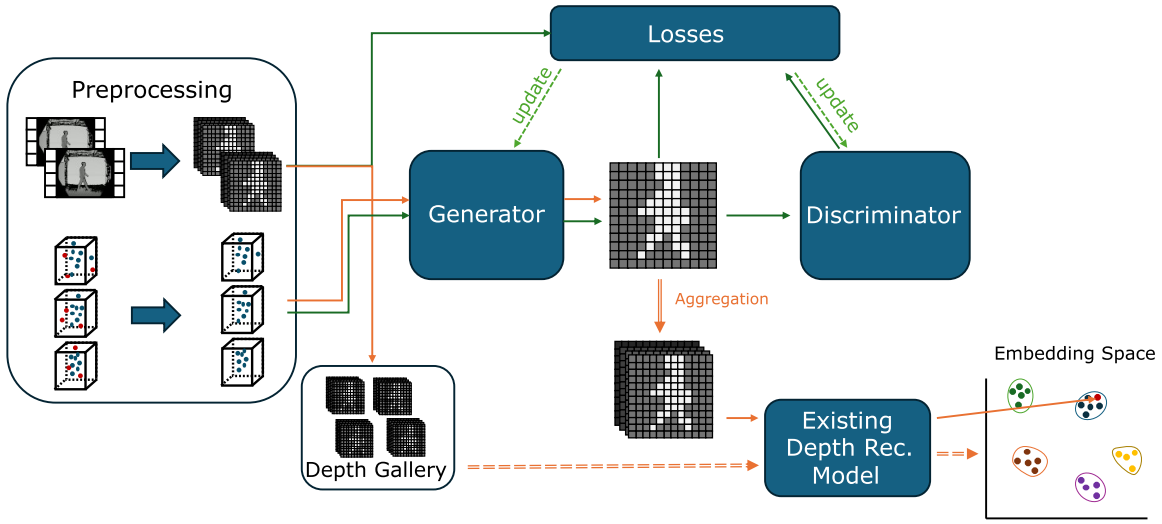



**Figure 4.1.:** Complete Model Pipeline. Dataflow during training is shown in green, and during inference in orange. The interrupted line towards the existing depth recognition model represents the mapping of the complete depth gallery into the embedding space.

The pipeline begins by preprocessing the radar point-cloud frames and the depth sequence frames. The central task for the depth frames is to extract silhouettes and produce a format compatible with the depth-based recognition model. This includes silhouette extraction, cropping, and resizing. For the radar point-cloud frames, preprocessing focuses only on denoising each frame to produce a cleaner representation of the walking subject. After preprocessing, both modalities are available in a training-ready paired dataset comprising radar point-cloud frames and corresponding depth-silhouette frames.

These frame pairs are then used for training of the GAN. At each step, the generator takes a radar point-cloud frame as input and generates a corresponding synthetic depth-silhouette frame. The discriminator, on the other hand, receives the same radar-derived condition and the generated and the real frame. The goal of the discriminator is to learn to distinguish the real silhouette frames from the generated silhouette frames under the given condition. On the other hand, the generator learns to produce silhouettes that are both realistic and consistent with the radar input.

Afterwards, the loss functions are evaluated using the real and synthetic silhouette and the discriminator output. It aims to combine multiple complementary losses. Reconstruction-

based loss for image similarity, adversarial-based loss for realism, and additional task-specific losses to improve overlap quality and recognition embedding similarity. The objective function of the discriminator, however, evaluates how well the discriminator can distinguish between the real and the generated silhouettes.

During inference, the generator is evaluated on radar point-cloud frames from a radar point-cloud sequence to generate the corresponding depth-silhouette frames. Afterwards, the depth silhouette frames are assembled in temporal order to create the complete synthetic depth silhouette sequence. This sequence can then be used as input to the depth-based gait recognition model to generate an embedding of the sequence.

For concrete identification, a comparison of this synthetic sequence embedding with embeddings of real depth-silhouette sequences of known identities is performed. The gallery sequence embedding is created using the same depth-based gait recognition model. The identity is then derived from the closest-gallery embedding, depending on the chosen comparison metric.

## 4.5. Data Preprocessing

The proposed approach requires the paired radar and depth recordings to be sufficiently cleaned and well-aligned. This is needed to properly train the generative adversary model. Because the adversary uses the dataset without identity labels or any supervision beyond recording-level pairing, the quality of the learned model depends even more on the consistency and precision of the training data. In the following subsection, the preprocessing steps applied to each sensor domain are described to generate a cleaned dataset suitable for training.

### 4.5.1. Radar Data

As stated in the Section 2.2, radar point cloud data is susceptible to artificial ghost points that do not correspond to physical objects in the observed scene. A widely used approach to denoise a radar point cloud is to apply DBSCAN [13], a density-based clustering algorithm, to identify the largest cluster and retrieve the associated data points [54, 6]. However, this can often yield denoised frames that are incorrect depending on the noise level. To address this, we propose another denoising pipeline. This pipeline comprises multiple stages: frame aggregation, duplicate removal, density-based clustering, and bounding-box-based reinsertion of duplicates.

A naive approach for denoising is to apply density-based clustering directly to individual radar frames. This can be insufficient. Single radar frames are sparse, with genuine object reflections only in a small number of points per frame. Applying density-based clustering under these conditions cannot reliably extract a cluster of valid points, as both real and noise point clusters exhibit similar low point densities. To address this, consecutive frames are aggregated before clustering. By incorporating the points of multiple frames, the number of

real points increases substantially. The number of aggregated frames is set to 7 to increase the time interval to a typical time interval for a single radar frame used in the literature [6].

When applying density-based clustering on the aggregated frames, a new concern arises. Spatially redundant points in neighboring frames become dominant. This leads to artificial inflation, particularly in persistent noise regions. Applying density-based clustering at this stage would result in many noise cluster, as the biggest cluster. Therefore, the remaining data would consist only of noise. To mitigate this issue, points at nearly identical positions in 3-dimensional space are replaced by a single representation, lowering the density of these regions.

After the deduplication step, DBSCAN can be successfully applied. It partitions the point cloud into clusters based on local density, retaining only the largest cluster for our purposes. This assumes that the dominant cluster corresponds to the primary object of interest, which has been shown to represent the human body [54, 6].

Afterwards, the axis-aligned bounding box of the retained cluster is computed. All previously removed duplicate points that fall into the bounding box are reintroduced to ensure legitimate points are recovered. Leading to the final result of the de-noising pipeline.

### 4.5.2. Depth Data

The depth sequence comprises frames depicting the entire scene, including the background. As motivated in Section 4.2, these are converted into silhouette depth frames. The corresponding silhouette extraction, cropping, and resizing follow the implementation of Todt et al. [41].

Silhouette extraction is performed by background subtraction using a static reference frame of the empty scene. Subtracting the reference from each depth frame isolates the foreground region corresponding to the moving subject. Afterwards, operations are applied to suppress pixel-level noise, retaining only large, coherent blobs corresponding to the subject's silhouette.

The silhouette sequence is then cropped and resized to a fixed resolution matching the input requirements of the gait recognition model. This eliminates the need for additional post-processing after generating the synthesized frames. To achieve this, bounding boxes are computed per frame using the white-pixel expansion of the silhouette. To avoid jitter across consecutive frames, bounding boxes are temporally smoothed using a sliding-window median filter. With these smoothed bounding boxes, the frames are cropped and resized to the target resolution.

After generating the cropped silhouette sequences, a new issue presents itself: the start and end frames of the sequence consist of empty frames. These represent frames in which the person was either not in the field of view yet or had already left it. However, these would provide false information during training for frame pairs in which radar point-cloud data are available, but no silhouette is present. Because this artifact is consistent throughout the

recordings, a simple omission of a fixed number of frames at the start and end is sufficient. This ensures that all retained frames contain a visible subject.

## 4.6. Model Architecture

As motivated in Section 4.2, the proposed architecture follows a conditional GAN architecture, comprising a generator and a discriminator. The generator produces synthetic depth-silhouette frames from radar point-cloud inputs. It is composed of an encoder that maps the radar point-cloud input to a fixed-size latent representation, and a decoder that maps this representation to a depth-silhouette frame. The discriminator then assesses the realism of the generated frames conditioned on the point-cloud encoding, thereby providing an adversarial training signal.

### 4.6.1. Encoder

The encoder is based on the PointNet architecture described in Section 2.2.1. It was chosen due to its successful application to radar point clouds for gait recognition [6] and image generation tasks on point cloud data [24]. While PointNet++ augments this architecture by capturing local neighborhood geometrical patterns, its use for radar point clouds is not justified: The low point densities typical of radar point clouds lead to local neighborhoods being poorly defined and therefore not worth analyzing. The encoder, consequently, builds on the base PointNet architecture, with three modifications to adapt it to this thesis' task: removing the input transformation, enriching the max-pooling output, and using mask-aware pooling for batched training. Figure 4.2 shows the resulting architecture, with each modification from the standard PointNet architecture visualized. The following paragraphs explain each modification in turn.

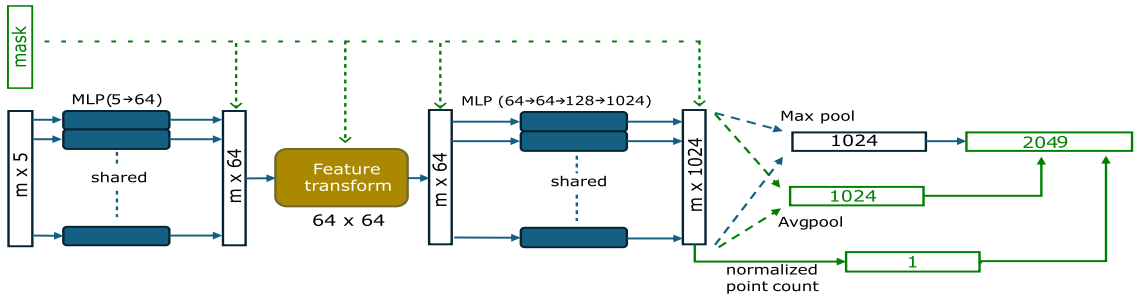



**Figure 4.2.:** Adapted PointNet Encoder Architecture. The green-highlighted parts are newly added, and the yellow-highlighted parts were modified from the original PointNet.

The first major change is removing the first transformation network, which maps the input to an invariant representation. The reason for this change lies in the preprocessing and in the difference between our task and that of the original PointNet application. In the paper, Qi et al. [28], only evaluate PointNet on classification and segmentation of objects

and scenes, which is expected to work regardless of the object's rotation and viewpoint. However, in this specific setting, we have a translation task in which the observed persons move only laterally along a fixed axis, so no geometric invariance is required after input centering during preprocessing. Additionally, this could harm potentially important spatially peripheral points representing arms or legs that are scaled towards the centroid of the data points during alignment, thereby lessening their influence on the encoding. Other radar point-cloud gait-recognition tasks use the PointNet architecture in its original form, applying the input coordinate transformation [3, 6, 49] while skipping the transformation of velocity values, as they do not satisfy the affine-invariance assumptions [6]. However, they also focus on a different task of identifying multiple potential subjects, regardless of the moving direction or positioning. This implies a need for geometric invariance and alignment that does not apply to our setting.

As a second change, the encoding is enriched beyond the max pooling present in the original PointNet architecture. The goal is to capture more of the potential information in the point cloud by addressing shortcomings of max pooling. This results in a new combined global encoding

$$z = [z_{max}; z_{avg}; c] \in \mathbb{R}^{2049}$$

where $z_{max}$ and $z_{avg}$ are the max- and average pooling of the 1024-dimensional feature vectors, and $c$ is a normalized point-count scalar. While max pooling provides useful insights into salient features, it completely discards information about how features are distributed across all points. For dense point clouds, this mainly removes redundant information. However, given the sparsity of radar point clouds, this distributional information is not redundant and might remove important details [49]. Average pooling aims to retain this information, thereby enriching the final encoding. The scalar c explicitly reintroduces point density, which is otherwise lost after global aggregation.

Because PointNet was originally designed for dense point clouds, a fixed sampling count is used as an upper bound for the input. However, for radar point clouds, determining an appropriate cut-off point is challenging due to their sparsity and varying point counts. This introduces issues because batched training requires a consistent input size across the samples. To achieve this, shorter inputs are padded with zero-value points to a fixed length. This, however, introduces the issue that padded points may influence the final encoding during training. To prevent this, the encoder accepts an auxiliary binary mask indicating which points are valid and which are padded. These are used throughout the encoder to reset the point-specific features of padded points, thereby removing their influence. For max pooling, padded entries are set to $-\infty$ prior to aggregation, for average pooling, they are excluded from the denominator, and the point-count scalar is derived directly from the mask. These adaptations are made to remove the influence of padded points. The choice of maximum point count and further details on masking are discussed in the experimental setup during evaluation, as they are relevant only because batches are used for training.

### 4.6.2. Decoder

The decoder maps the encoder output $z \in \mathbb{R}^{2049}$ from the latent space to a single-channel depth image of size $64 \times 64$ via four upsampling stages, progressing spatially from $4 \times 4$ to $64 \times 64$. A convolutional upsampling decoder is a well-established choice for this class of image generation tasks [24, 14], progressively reconstructing spatial structure from a compact latent representation through a sequence of upsampling stages. The architecture shares the multi-scale progressive structure of U-Net-style decoders [30] but does not employ explicit skip connections from the encoder, as no paired-image encoder exists in this pipeline to enable them.

A standard upsampling decoder conditions on the latent code only at the bottleneck, which limits the influence of the encoding on later, higher-resolution stages of generation. In a U-Net-style architecture, this issue can be mitigated by skip connections that deliver additional information from the encoder stages to the parallel decoder stages. For our approach, FiLM conditioning, as described in Section 2.5.3, is therefore applied at each upsampling stage with stage-specific learned parameters. This enables the global encoding to perform coarse-to-fine conditional control throughout the decoding process at each stage. The goal of this addition for this approach is to substitute the encoder context that skip connections would otherwise provide to the decoder stages. The related Adaptive Instance Normalization (AdaIN) [18] operates on the same principle of feature-wise affine modulation but is designed specifically for style transfer in image-to-image translation. It can be seen as a specialization of FiLM. There, the conditioning signal encodes the appearance statistic of a style image. Since, in our task, the encoder output is a semantic embedding of the radar point cloud rather than a style representation, FiLM is the more appropriate mechanism.

The full architecture is shown in Figure 4.3

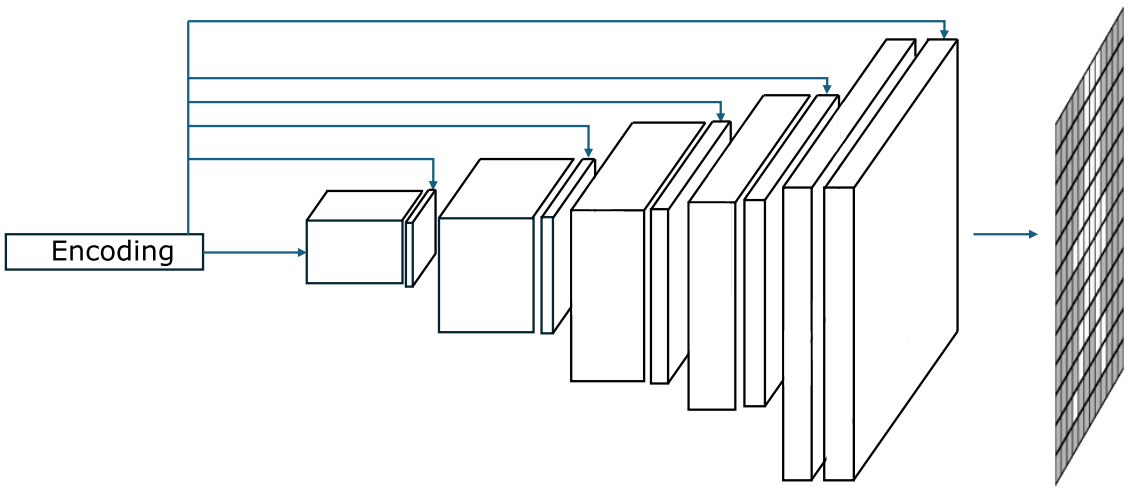



**Figure 4.3.:** Decoder Architecture

### 4.6.3. Discriminator

The Discriminator follows the PatchGAN design [20], as described in Section 2.5.1. It produces a spatial map of local realism scores rather than a single image-level decision. This is especially relevant when errors are spatially localised and has proven to be a suitable design for GAN tasks with image-like outputs [24]. The base consists of four convolutional blocks that progressively downsample the input depth map into a spatial feature volume. In the final convolution, this feature volume is mapped to a dense patch of logit values. The full architecture is shown in Figure 4.4.

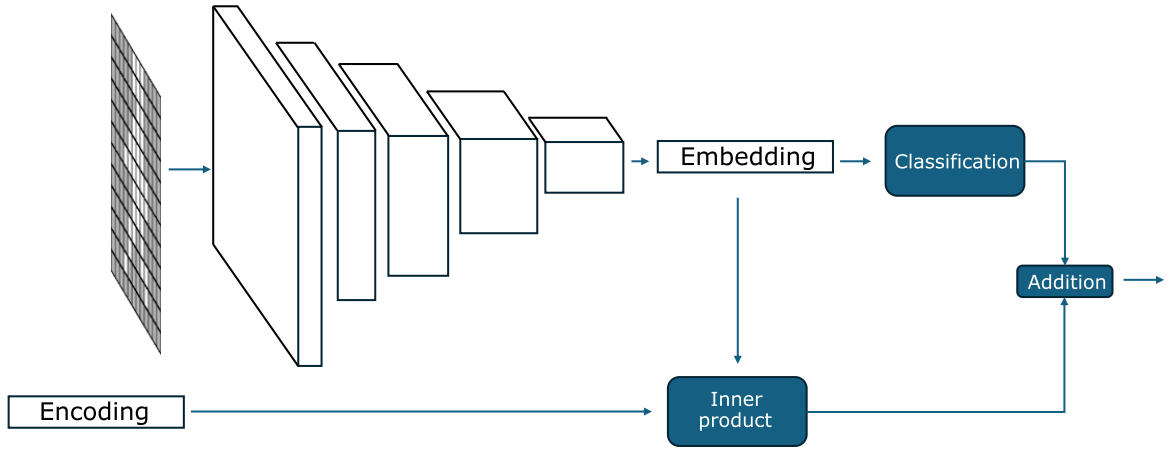



**Figure 4.4.:** Discriminator Architecture

To condition the discriminator on the point cloud, the projection mechanism proposed by Miyato and Koyama [26] is adopted in preference to simpler alternatives. Two such approaches are to concatenate the raw point cloud directly or to concatenate $z$ to the discriminator's input.

Conditioning a convolutional discriminator on the raw point cloud is impractical. Point clouds are variable in size and unordered. Incorporating them directly would require an additional processing branch, such as the original point cloud encoding. This would introduce substantial additional complexity. Since the encoder already produces a compact fixed-size embedding $z$ that captures the task-relevant structure of the point cloud. Conditioning on $z$ avoids the redundancy of a separate encoding.

This introduces the option of concatenating the condition to the discriminator's input or to a hidden layer as a next common alternative. Miyato and Koyama [26] state, however, that this is probabilistically unprincipled. It introduces an arbitrary function of the image and the condition, without grounding it in the assumed conditional distribution. Under the assumption that the conditional distribution is unimodal, the optimal discriminator can be described by the Equation 2.1 introduced in Section 2.5.1. Our approach is trained on a frame-to-frame paired basis with deterministic generation. While this does not prove

unimodality of the true conditional distribution, it makes projection-based conditioning a plausible and appropriate design choice.

In the original formulation, global sum pooling collapses the spatial feature map to a single vector. Only afterwards is the inner product computed, producing a scalar conditioning term suitable for a global discriminator. However, because we use a PatchGAN discriminator without the global pooling step, the inner product is computed independently at each spatial location. This yields a per-patch condition term that is added directly to the patch logits. Thereby adapting the original concept of condition projection of Miyato and Koyama [26] to the PatchGAN architecture. This results in an output map reflecting local realism conditioned on the input image and the global point cloud encoding.

## 4.7. Loss Function

The total training objectives contain a separate generator and discriminator loss. The discriminator loss follows the LSGAN definition (Section 2.5.2). The generator loss, on the other hand, is comprised of an adversarial term, a reconstruction term, an overlap term, and a gait term:

$$\mathcal{L} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ovlp}}\mathcal{L}_{\text{ovlp}} + \lambda_{\text{gait}}\mathcal{L}_{\text{gait}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} \tag{4.1}$$

Each term targets a distinct aspect of the generation task. $\mathcal{L}_{\text{adv}}$ drives local structural realism, $\mathcal{L}_{\text{rec}}$ provides global geometry, $\mathcal{L}_{\text{ovlp}}$ enforces silhouette shape, and $\mathcal{L}_{\text{gait}}$ aims to incorporate performance on downstream task. Finally, $\mathcal{L}_{\text{ortho}}$ represents a loss that regularizes the feature transformation network of the encoder to stay close to the identity. All other lambdas are derived by hyperparameter optimization described in Section 4.8. The following subsections motivate and define each term.

### 4.7.1. Adversarial Loss

The goal of the adversarial loss is to capture the generator's objective in the original GAN definition. However, we explicitly chose the least-squares GAN (LSGAN) [23] concept. It replaces the original binary cross-entropy objective with a least-squares objective that penalizes even samples that are on the correct side of the decision boundary but far from it. Otherwise, the objective is susceptible to vanishing gradient when the discriminator becomes too confident. The adversarial loss, therefore, is defined as follows:

$$\mathcal{L}_{\text{adv}}(D, \hat{y}, z) = \frac{1}{2}\text{mean}((D(\hat{y}, z) - 1)^2)$$

where $D$ represents the discriminator, $\hat{y}$ is the generated depth silhouette frame, and $z$ represents the encoding of the original point cloud on which the discriminator is conditioned. The mean is required because the output of $D$ is not a single value but a spatial map of local realism scores. This is due to the use of a PatchGAN discriminator, $D$.

### 4.7.2. Reconstruction Loss

The adversarial loss alone does not constrain the global structure. The generator can still satisfy the discriminator while producing outputs that are locally sharp but globally inconsistent. Especially with sparse input, where the generator is underconstrained and therefore more prone to hallucinated structures. Isola et al. [20] propose the addition of an L1 loss as a reconstruction loss to guide the global geometry for image generation tasks. It penalizes pixel-wise deviation and, when combined with the adversarial loss, produces outputs that are more globally consistent and locally sharper than those obtained with the adversarial loss alone. L1 is preferred over L2 because L1 encourages less blurring [20]. The L1-based reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}}(y, \hat{y}) = ||y - \hat{y}||_1$$

### 4.7.3. Overlap-based Loss

The reconstruction loss acts on all pixels equally. For silhouette generation, the relevant task can be viewed as segmenting the image to extract the foreground region. Pixel-wise error, however, does not explicitly optimize the spatial overlap between a generated and ground-truth segment. For segmentation tasks, losses that operate on an overlap basis are common [40, 31]. Two of those are the Dice coefficient and the Tversky index. Both of them describe the ratio of intersection to overlap. While Dice assigns equal weight to false negatives and false positives, Tversky introduces an explicit weighting of both. This allows more fine-tuning for the task at hand and can also be interpreted as Dice if false negatives and false positives are weighted equally. For this reason, Tversky is chosen over Dice because of its variability.

In silhouette generation, false negatives correspond to missing parts of the silhouette, whereas false positives correspond to additional generated parts. This allows individual tuning depending on problem areas. Although Tversky is defined on explicit binary fields, it can be applied to continuous values. This also avoids non-differentiable threshold variants, which would interrupt gradient flow. The following defines the overlap loss:

$$\mathcal{L}_{\text{ovlp}}(y, \hat{y}) = 1 - \frac{\sum_i y_i \cdot \hat{y}_i}{\sum_i y_i \cdot \hat{y}_i + \alpha \cdot \sum_i (1 - y_i) \cdot \hat{y}_i + \beta \sum_i y_i \cdot (1 - \hat{y}_i)}$$

where $\alpha$ and $\beta$ represent the weights for false-positive and false-negatives, which are parameters that are fine-tuned.

### 4.7.4. Gait embedding-based Loss

The adversarial, reconstruction, and overlap losses are defined to assess realism, pixel-level accuracy, and overlap accuracy. However, none of them directly optimizes for the downstream task of gait recognition. The goal is to produce outputs that the pre-trained

gait-recognition model can correctly identify. Generated silhouettes can perform well on all three losses, while the corresponding sequence is difficult for the recognizer to correctly identify.

A gait embedding loss is introduced to address the disconnect between the downstream task and the objective that is optimized. The loss aims to minimize the distance between the embeddings of the generated silhouette and the ground-truth silhouette in the feature space of the recognition model. The loss is defined as the mean squared error between the real and fake embedding:

$$\mathcal{L}_{\text{gait}}(R, y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (R(y)_i - R(\hat{y})_i)^2$$

where $R$ describes the recognition model, $y$ the real depth silhouette, $\hat{y}$ the generated silhouette and $n$ the length of the embedding vector However, a limitation of this approach is the GAN's frame-wise generation. The gait recognition model is trained on silhouette sequences rather than on single frames. Therefore, a single frame may produce an embedding that is not discriminative enough to produce a helpful signal. Consequently, the gait loss should be treated as an auxiliary signal.

### 4.7.5. Feature Transformation Orthogonality Loss

The goal of this loss is to regularize the feature transformation network to stay close to an orthogonal matrix. It is proposed by Qi et al. [28] for the original PointNet architecture and adapted here. It is motivated by the high dimension of the transformation matrix and the resulting difficulties with its optimization.

$$\mathcal{L}_{\text{ortho}}(A) = ||I - AA^T||_F^2$$

where $A$ is the feature transformation matrix, $I$ is the identity matrix and $|| \cdot ||_F$ is the Frobenius norm. An orthogonal transformation matrix is desired because it preserves all information in the input. Qi et al. [28] additionally also propose that $\lambda_{\text{ortho}} = 0.001$ which is therefore also applied here.

## 4.8. Hyperparameter Optimization

The lambdas defined in Equation 4.1, $\alpha$, $\beta$, and learning rates for generator and discriminator are optimized using Optuna [1], with the exception of $\lambda_{\text{ortho}}$, which is already suggested by its source. To mitigate the influence of a single random partition, hyperparameter tuning was performed across 4 random data splits. The resulting hyperparameter estimates were then averaged. This average was used as a common configuration across all experiments. While this creates a consistent, less split-dependent setup, it may not reflect the data-split-specific optimum for each split. Additionally, note that the tuning and evaluation partitions are not guaranteed to be disjoint at the subject level, which might introduce an optimistic bias. This will later be further discussed in Sections 6.4.

# 5. Evaluation

This chapter evaluates the proposed radar-to-depth pipeline with respect to the central question of this thesis. To answer this question, Section 5.1 defines the evaluation protocol, including the use of data, the training setup, and training details. Section 5.2 investigates the primary identification results. Section 5.3 investigates ablation studies on temporal context, loss functions, and architectural components. Finally, Section 5.4 analyzes model behavior in more detail rather than overall performance.

## 5.1. Evaluation Protocol

The evaluation investigates whether radar recordings can be translated into synthetic depth-silhouette sequences compatible with a depth-based gait-recognition model, thereby enabling downstream person identification. More specifically, the goal is not to assess direct radar-based gait recognition, but to examine whether the proposed radar-to-depth generator preserves identity-related gait characteristics well enough for a pre-trained depth silhouette recognizer to operate on the generated sequence.

To achieve this, the evaluation is framed as an identification problem in a gallery-probe setting. Real depth recordings are used to construct a gallery in the recognizer's native input domain, while probe samples are derived from radar recordings, which are then translated into synthetic depth silhouette recordings. Recognition performance is measured in the embedding space of the existing depth-based gait recognition model.

### 5.1.1. Dataset and Preprocessing

As a dataset, MultiGait [41] was chosen as it explicitly contains paired radar and depth recordings of different subjects. Additionally, it provides a set of attributes for each existing subject. It contains recordings of 189 subjects. For each subject, up to 3 recording sessions are available, although some subjects are represented in only 1 or 2 sessions. Each session contains up to 40 recordings, each consisting of one radar point-cloud recording and one corresponding depth recording with a synchronized start time. The average recording duration is approximately 4 seconds.

Radar and depth are paired at the recording level, not the frame level. This distinction is important for the translation setup, as paired data provide synchronized walking sequences but do not perfectly match radar and depth frames.

Preprocessing is performed independently for each recording and does not rely on any parameters estimated from the complete dataset. For the radar point clouds, they are first partitioned into frames based on timestamp gaps, with frame borders defined by a timestamp jump of at least 30 milliseconds. This value was chosen by inspecting the timestamp jumps in the radar point cloud data, as new frames were clearly signaled by jumps in timestamps from $\sim 34\,ms$. To allow some margin, a lower value was chosen. This does not cause any issues, as timestamps between radar points within a single frame are less than $1\,ms$. The resulting radar point-cloud frames are retained as the basic unit for alignment. They are explicitly not aggregated into disjoint multi-frame segments before alignment. The denoising described in Section 4.5 is then applied to these frame-split point cloud recordings.

For the depth recordings, silhouettes are extracted from the raw depth recordings, cropped, and resized to the target resolution of $64 \times 64$ as described in Section 4.5. This results in empty frames at the start and end of the resulting recordings. Therefore, the first and last 0.5 seconds of each recording are removed, since these regions have been shown to mainly contain empty frames before the subject enters or after the subject leaves the relevant field of view. Because most recording pairs differ in length, the end time is determined by the shorter recording. However, this is on average only 0.014 seconds.

The original set of depth recordings is available at 60 fps. After the preprocessing, it is downsampled to 30 fps. On the other hand, the radar point cloud is available at an effective frame rate of 29.4 fps. Nevertheless, radar and depth frames are matched using an index-based approach with synchronized start alignment. Because the recordings are short, the resulting temporal drift remains small, amounting to approximately 1.8 frames at the end of the recording.

Although alignment is performed at the level of individual frames, a single radar frame is typically too sparse to be used directly as model input. Therefore, temporal aggregation is applied when creating single samples. For a radar frame at index $t$, the model input is constructed by combining the corresponding frame with its local temporal neighborhood. For the given frame length, this value was set to 7 to correspond to existing frame lengths in related work [6], resulting in a frame window of $[t-3, \ldots, t, \ldots, t+3]$. This results in each training sample using a temporally aggregated radar point cloud centered on a single aligned frame index. On the other hand, the corresponding depth silhouette at that index remains the same. The aggregation is therefore part of the input representation, not a separate preprocessing step that alters the alignment procedure itself.

The cropped depth silhouettes are already represented in the range $[0, 1]$ after the preprocessing and are not subject to any additional normalization. Radar point clouds are centered to match the spatial representation of the depth silhouettes. Velocity values are also centered, since they may otherwise encode global positioning effects. The timestamp feature is normalized to the timestamp of the center index and then scaled by the timestamp step size, which was used to split the point cloud into separate frames. This results in them representing roughly the relative temporal position in the global dataset. All normalization is performed on a per-sample, per-frame basis. No global dataset statistics, such as global mean or variances, are used.

### 5.1.2. Split Construction and Session Selection

The dataset is split at the subject level into 80% for training and 20% for testing, using a fixed random seed. The resulting split is stored and reused across experiments to preserve reproducibility and comparability. Since the split is subject-disjoint, no identity appears in both training and test data. Additionally, four more splits are generated and saved to examine the variance of the trained model across different splits.

To avoid overrepresentation of subjects with multiple sessions, exactly one session is selected at random per subject. This session selection is performed once, using a fixed seed, to ensure consistency across experiments. As a result, each subject contributes equally at the recording level, and the evaluation protocol remains reproducible across model variants.

### 5.1.3. Training Setup

The 80% training split is used to train the radar-to-depth GAN model proposed in Section 4.6. The paired radar and depth recordings are converted into a frame-wise training set, as described in Section 5.1.1. These frame pairs are shuffled before training to reduce sequential correlation. No identity labels or other semantic annotations are used during training.

The GAN model is trained using an Adam optimizer [22]. The generator and the discriminator are updated at a 1:1 ratio without warm-up, and with one-sided label smoothing for the discriminator to mitigate the discriminator's overconfidence. The concrete values used for training can be seen in Table A.2 in the appendix.

Two training settings are considered for the GAN model: short-run and long-run. The short runs are used as the main evaluation models, since most of the optimization is already achieved after 100 epochs. The long run is used to assess the potential for improvement after 100 epochs. The long-run model is trained for 800 epochs. The final model in each case corresponds to the last checkpoint. The exact learning-rate scheduler used in the long run reduces the learning rate by 0.5 every 100 epochs since gains beyond 100 epochs remained marginal otherwise. For both, the batch size is 800 frame-wise training samples, each comprising one target depth silhouette and the aggregated radar frames, with frame size 7.

The depth-based gait recognition model is trained separately on the same subject-disjoint training split and is kept fixed. Consequently, the final evaluation is performed on identities that are unseen to both the radar-to-depth GAN model and the recognition model. At the same time, the evaluation remains a within-dataset protocol, since both components are trained on the same training partition. The reported results, therefore, assess whether the generated synthetic depth sequences are independently compatible with a frozen depth recognizer under subject-disjoint generalization, rather than transferred to an independently trained external recognition model.

### 5.1.4. Recognition Protocol

Final performance is evaluated exclusively on the held-out test split. For each test subject, 20 recordings are selected to construct the gallery. The corresponding real depth recordings are processed by the frozen depth-based gait recognition model to obtain gallery embeddings.

The remaining recordings of the same test subject are used as probe samples. In the experiments, only the radar point-cloud recordings are used as probe input. Each radar recording is translated frame by frame into synthetic depth silhouettes using the trained generator. The generated depth frames are then reassembled into a synthetic depth-silhouette sequence, which is passed to the same gait recognition model to compute the probe embedding. This follows the overall model pipeline.

Identification is performed by retrieval in the embedding space of the gait recognition model using cosine similarity. For each probe embedding, the closest gallery embedding is retrieved, and the identity associated with this gallery sample is used as the predicted identity.

### 5.1.5. Metrics

Recognition performance is reported using rank-based identification metrics. The primary metric is Rank-1 accuracy, which measures the proportion of probe samples for which the correct identity is retrieved as the nearest gallery match. In addition, Rank-5 accuracy is reported to assess whether the correct identity appears among the five closest gallery candidates. In cases in which the same identity is present under the five nearest galleries, the duplicates are removed such that the five nearest gallery embeddings are from distinct identities.

Where appropriate, recognition results are complemented by image-level similarity scores between the generated and real depth silhouettes, computed using the Structural Similarity Index Measure (SSIM) [48]. However, these image-level metrics are treated as auxiliary measures of reconstruction quality rather than as direct indicators of identification performance.

### 5.1.6. Reference Systems and Baselines

The main reference system is the corresponding real-depth recognition setting, in which real depth recordings are used for both gallery and probe construction. This reference reflects the performance of the frozen gait recognizer and serves as an upper bound in the present evaluation protocol.

The primary system of interest is the radar-to-depth pipeline, in which real depth recordings define the gallery and synthetic depth sequences generated from radar define the probes. In addition, the chance level is reported as a trivial lower bound.

### 5.1.7. Implementation Details

All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU. The software environment was based on Python 3.12.12 and PyTorch 2.5.1 with CUDA 12.1 and cuDNN 9.1.0.

## 5.2. Primary Results

This section presents the main empirical result of the thesis. The central question is whether the proposed radar-to-depth pipeline produces synthetic depth sequences that are sufficiently compatible with the frozen depth-silhouette-based gait recognition model to enable downstream identification. The overall performance and the variance across different dataset splits are investigated.

### 5.2.1. Hypothesis

To investigate the main contribution of this thesis, the following hypothesis is evaluated:

**H1** The proposed radar-to-depth generator enables downstream identification above the chance level with an existing depth-silhouette-based gait recognition model.

This hypothesis is evaluated from two complementary perspectives. First, recognition performance is assessed across five subject-disjoint short-run training splits to assess whether the effect is robust under a fixed training budget. Second, an extended training run on the first canonical split is used to assess the additional performance that can be obtained beyond the short-run model.

### 5.2.2. Experiment

This experiment follows the evaluation protocol defined in Section 5.1. The primary analysis is based on five different dataset splits, created as described in Section 5.1.2. For each data split, a training run with a fixed length of 100 epochs is done. Each split contains 38 subjects in the test data. This results in a trivial chance level of 0.0263 for Rank-1 accuracy and 0.1316 for Rank-5 accuracy. This part of the experiment assesses whether the proposed radar-to-depth pipeline yields robust identification performance above the chance level across different splits. Furthermore, it provides insight into the variance of the trained model and its dependence on the training and testing subjects.

In addition, one extended training run is conducted on the first split for 800 epochs using the learning-rate schedule defined in the implementation details. This long-run experiment is included to estimate the performance potential beyond the short-run models.

**Table 5.1.:** Recognition performance across the five short-run splits. Reported are Rank-1 and Rank-5 identification accuracies for real depth probe sequences (GT) and generated probe sequences (GAN).

| Split | GT R@1 | GT R@5 | GAN R@1 | GAN R@5 |
|---|---|---|---|---|
| Split 1 | 0.9997 | 1.0000 | 0.0666 | 0.2232 |
| Split 2 | 0.9935 | 1.0000 | 0.0440 | 0.1884 |
| Split 3 | 0.9896 | 1.0000 | 0.0636 | 0.2488 |
| Split 4 | 0.9986 | 1.0000 | 0.0451 | 0.1836 |
| Split 5 | 0.9993 | 1.0000 | 0.0635 | 0.2182 |
| Mean ± SD | 0.9943 ± 0.0035 | 1.0000 ± 0.0000 | 0.0566 ± 0.0110 | 0.2124 ± 0.0268 |

**Table 5.2.:** Short-run versus long-run performance on the first split.

| Run | GAN R@1 | GAN R@5 |
|---|---|---|
| Short (100 epochs) | 0.0666 | 0.2232 |
| Long (800 epochs) | 0.0750 | 0.2650 |

For both, recognition performance is measured using the gallery-probe identification protocol and the rank-based metrics defined in Section 5.1.5. Results are reported per split, together with an aggregate summary across the five short-run splits. The comparison between the short-run and long-run model on the first split is reported separately.

### 5.2.3. Results

Table 5.1 summarizes the recognition performance across the five short-run splits. Recognition on real depth silhouettes is nearly perfect with an average Rank-1 accuracy of 0.9943 and a Rank-5 accuracy of 1.0. In contrast, recognition on generated silhouettes remains substantially lower but consistently above the chance level.

Averaged across the five short-run splits, the proposed radar-to-depth pipeline achieves a GAN Rank-1 accuracy of 0.056 (±0.0110) and a Rank-5 accuracy of 0.2124 (±0.0268). Performance on split 1 is at its highest with a Rank-1 accuracy of 0.0666 and at its lowest for split 2 with a Rank-1 accuracy of 0.0440. However, the highest Rank-5 accuracy is not achieved on split 1 but on split 3 with 0.2488, while the lowest is achieved for split 4 with 0.1836

Table 5.2 compares the short-run and the long-run models on the first data split. Extending the training improves the Rank-1 accuracy from 0.0666 to 0.0750 and the Rank-5 accuracy from 0.2232 to 0.2650.

For transparency, Appendix Table A.1 reports subject-level GAN Rank-1 and Rank-5 accuracies across all five splits. This overview allows repeated subjects to be compared directly

across different test-set realizations. However, in this section, the overall performance is inspected.

### 5.2.4. Findings

The results support H1. Across all five subject-disjoint short-run splits, the proposed radar-to-depth pipeline enables downstream identification above chance level with the frozen depth-silhouette-based gait recognition model. This indicates that the generated silhouettes preserve some identity-related gait information. However, at the same time, the gap to the real-depth reference remains very large. While the performance is better than chance, it is only by a small absolute margin and is far from that of a good recognition model.

Since recognition on real depth silhouettes is nearly perfect under the same protocol, the observed limitation cannot be attributed to the gallery-probe setup in itself. Additionally, insufficient gait-recognition performance cannot be considered as a root cause. Instead, the main limitation must arise from incomplete preservation of discriminative features during translation.

Besides that, another pattern can be observed. An increase or decrease in Rank-1 accuracy does not imply a corresponding change in Rank-5 accuracy. While Rank-1 accuracy is reduced by 0.003 when comparing split 1 and split 3, the Rank-5 accuracy is increased by 0.0256. On the other hand, split 3 has a similar Rank-1 accuracy as split 5, but the Rank-5 accuracy of split 5 is lower by 0.0306. This shows that Rank-1 and Rank-5 accuracies are not directly related because they correspond to different retrieval approaches.

Furthermore, the variance in accuracy must be reported as it is apparent. While the standard deviation of 0.011 does not seem large, it is when compared to the concrete values. Given a mean of 0.0566, this deviation presents a major part of the overall performance. This shows that the concrete performance of the radar-to-depth pipeline depends on the data split used and can vary noticeably.

Finally, the split-wise averages inspected here conceal substantial variation at the subject level that can be observed in Table A.1 in the Appendix. While some subjects achieve comparatively high recognition rates, many identities remain difficult to recover from the generated depth silhouettes. This observation motivates the subject-wise analysis in the subsequent failure analysis in Section 5.4.1, which investigates it in more detail.

## 5.3. Ablation Studies

While the primary result shows whether the recognition pipeline works, it does not explain the influence of single components of the system. The following ablation studies, therefore, examine the contribution of major design decisions. Specifically, they investigate the effects of the input temporal context, the loss function, and the architectural components.

**Table 5.3.:** Temporal-context ablations on the first split.

| Variant | GAN R@1 | Δ R@1 | GAN R@5 | Δ R@5 |
|---|---|---|---|---|
| Baseline | 0.0666 | – | 0.2232 | – |
| bigger window | 0.0454 | -0.0212 | 0.2306 | +0.0074 |
| smaller window | 0.0481 | -0.0185 | 0.1614 | -0.0618 |
| without timestamp | 0.0653 | -0.0013 | 0.2367 | +0.0135 |

### 5.3.1. Temporal Context

In Section 4.5, the choice of the window size was justified by the length of radar frames, as seen in the literature. However, neither here nor in the referenced research was an argument given for this size. Therefore, the influence of larger and smaller window sizes is inspected. Additionally, the addition of timestamps is investigated, as the overall aggregation already implies some form of frame affiliation. This leads to the following hypotheses.

**H2** A temporally aggregated radar context improves translation quality and downstream recognition performance. Too little temporal context is expected to underrepresent motion cues, whereas large context windows may weaken alignment or fail to provide significant improvement.

**H3** The addition of timestamps to the input only provides a slight improvement.

#### 5.3.1.1. Experiment

To analyze the contributions and effects of the temporal context, training runs are conducted by varying the size of the radar aggregation window while keeping all other components fixed. Two short training runs are performed on the first data split and compared with the original model trained on the same split. One is trained with a cumulative window size of 3, while the other is trained with a cumulative window size of 11. Additionally, one model is trained with the timestamp attributes removed from the input. All models are evaluated using the primary gallery-probe protocol. The results are compared with the original short-run model on the same data split, using Rank-1 and Rank-5 accuracy.

#### 5.3.1.2. Results

Table 5.3 summarizes the temporal-context ablations on the first split. The baseline short-run model is introduced in Section 5.2 and the Rank-1 and Rank-5 accuracies are displayed here again. Additionally, the concrete accuracies of the variant models and their deltas relative to the baseline short-run model are shown.

Reducing the aggregation window reduces performance to 0.0481 at Rank-1 and 0.1614 at Rank-5. Increasing the window size also reduces Rank-1 accuracy to 0.0454, while Rank-5

changes only slightly to 0.2306. However, removing the timestamp feature leaves Rank-1 almost unchanged at 0.0653 and slightly increases Rank-5 to 0.2367.

#### 5.3.1.3. Findings

The temporal-context ablations indicate that the chosen aggregation window represents a meaningful trade-off. Reducing the temporal context clearly harms downstream recognition, especially in Rank-5. This suggests that a limited local radar context is insufficient to reliably preserve identity-relevant gait information. On the other hand, increasing the window size does not yield consistent improvements. Although Rank-5 remains close to the baseline, Rank-1 decreases noticeably, which suggests that a larger temporal context may weaken the alignment between the radar input and the target depth frame. The larger the window size, the more similar the two consecutive aggregated frames. This might impair the translation of the aggregated frame into a depth silhouette. These insights support H2 and, therefore, the choice of window size.

In contrast, removing the timestamp feature has almost no effect on the downstream Rank-1 accuracy. However, Rank-5 accuracy even improves slightly. This suggests that timestamp information is not essential for preserving identity-related cues. The approach of constructing the aggregated input by appending the point cloud data frame-by-frame might already give sufficient information about the corresponding original frame of each point in the input. However, timestamps may still be relevant to motion direction, as will be discussed in Section 5.4.3.

### 5.3.2. Objective Design

While Section 4.7 motivated the choice of the loss terms, it remains unclear how much influence each has on the downstream identification. Preliminary observations during training suggest that the different terms may not contribute equally. The weight $\lambda_{\text{adv}}$ received a comparably small weight during hyperparameter tuning, while the recognition loss remained relatively stable and low. Therefore, the following hypotheses are formulated:

**H4** Adding the gait-recognition loss does not yield a measurable improvement in downstream recognition performance relative to the baseline model without this loss term.

**H5** The adversarial loss does not provide a measurable benefit for downstream recognition performance; removing it is not expected to reduce performance relative to the full model.

#### 5.3.2.1. Experiment

To analyze the contribution of different training objective terms, the full model is compared with variants that remove or replace individual loss terms. In particular, the adversarial

**Table 5.4.:** Objective-design ablations on the first split.

| Variant | R@1 | Δ R@1 | R@5 | Δ R@5 |
|---|---|---|---|---|
| Baseline | 0.0666 | – | 0.2232 | – |
| no adversarial loss | 0.0703 | +0.0037 | 0.2449 | +0.0217 |
| no gait loss | 0.0663 | -0.0003 | 0.1855 | -0.0377 |

loss, the gait-related loss, and the overlap-based loss term are examined through multiple ablations.

Four variants are trained with the same training settings as the original short-run model on the first data split. Both variants investigate the simple removal of each loss term. The comparison examines whether removing a given loss term degrades downstream recognition.

#### 5.3.2.2. Results

Table 5.4 reports the objective-design ablations on the first split. Removing the adversarial loss yields a GAN Rank-1 accuracy of 0.0703 and a GAN Rank-5 accuracy of 0.2449. This shows an increase in comparison to the baseline. On the other hand, removing the gait-related loss results in a GAN Rank-1 accuracy of 0.0663, which is nearly unchanged relative to the baseline, while Rank-5 accuracy decreases to 0.1855.

#### 5.3.2.3. Findings

The available objective-design ablations suggest that the different loss terms do not contribute equally to the downstream recognition. In the current setting, removing the adversarial loss does not degrade performance and even yields a slight improvement on the evaluated split. This insight supports H5. It indicates that adversarial training is not required for downstream accuracy under the current training protocol. This implies that an adversarial generative approach may not be appropriate for the defined problem and may actually slightly improve the result.

The gait-related loss shows a more differentiated pattern. Its removal leaves Rank-1 accuracy almost unchanged, but reduces the Rank-5 accuracy noticeably. This suggests that the gait-related loss may not be critical for exact top-rank identification, but may still support a broader preservation of identity-relevant information. Accordingly, the current results do not support claim H4, as it improves Rank-5 accuracy by 0.037, which is clearly measurable.

**Table 5.5.:** Architecture-design ablations on the first split.

| Variant | GAN R@1 | Δ R@1 | GAN R@5 | Δ R@5 |
|---|---|---|---|---|
| Baseline | 0.0666 | – | 0.2232 | – |
| with input transform | 0.0576 | -0.0090 | 0.2404 | +0.0172 |
| no feature transform | 0.0674 | +0.0008 | 0.2021 | -0.0211 |
| no FiLM | 0.0328 | -0.0338 | 0.1514 | -0.0718 |

### 5.3.3. Architecture Design

Whereas in the other parts, the reasoning for the architecture choices was motivated at a theoretical level or by related work. However, the exact influence on the downstream identification performance of each part is unknown. It is therefore necessary to determine whether all included components are beneficial, or whether some parts are unnecessary or even counterproductive in the present setting. This leads to the following hypotheses:

**H6** Reintroducing the input transformation network does not improve downstream recognition performance and may reduce it relative to the baseline model.

**H7** Removing the feature transformation module reduces downstream recognition performance relative to the baseline model.

**H8** Removing FiLM-based conditioning in the decoder reduces downstream recognition performance relative to the baseline model.

#### 5.3.3.1. Experiment

The architectural ablations are performed by removing or modifying individual network components while preserving the rest of the pipeline. This includes the input transformation network (H6), the feature transformation network (H7), and the conditioning FiLM layers (H8).

Each variant is again trained and evaluated under the same short-run protocol on the first data split. Performance is compared against the full model using the primary recognition metrics. These are used to determine whether architectural changes primarily affect visual reconstruction, downstream identification, or both.

#### 5.3.3.2. Results

Table 5.5 summarizes the architectural ablations on the first split. Reintroducing the input transformation network in the encoder reduces Rank-1 accuracy to 0.0576 and increases Rank-5 accuracy to 0.2404. Removing the feature transformation network yields a Rank-1 accuracy of 0.0674, nearly identical to the baseline, and a decrease to 0.2021 in Rank-5

accuracy. Removing the FiLM layers reduces Rank-1 accuracy to 0.0328 and Rank-5 accuracy to 0.1514.

#### 5.3.3.3. Findings

Among architectural ablations, FiLM-based conditioning has the clearest positive effect. Removing FiLM substantially degrades both Rank-1 and Rank-5 accuracies. This indicates that decoder-side incremental conditioning is an important change relative to single conditioning at the bottleneck and supports H8.

Reintroducing the input transformation network does not yield a clear overall benefit. While Rank-5 increases slightly, Rank-1 decreases relative to the baseline. This suggests that the additional transformation does not improve the exact downstream identification. Therefore, the increase in Rank-5 accuracy can not justify the decrease in Rank-1 accuracy. However, it cannot be determined that H6 is fully supported.

Removing the feature transformation network leaves Rank-1 nearly unchanged but reduces Rank-5 accuracy. This indicates that the component primarily contributes to retrieval robustness rather than to exact top-rank identification. However, this shows that preserving the feature transformation network is beneficial to the downstream task. Therefore, H7 can not be confidently supported.

## 5.4. Model Behavior Analysis

The ablation studies identify the contributions of controlled design choices, but they do not fully explain the model's remaining limitations. While they provide insight that performance varies substantially across subjects, it is not investigated whether there is a possible explanation for this. They also do not assess whether image-level similarity improves recognition performance on the downstream task. Additionally, the seemingly good representation of the walking direction is also not investigated. The following analyses therefore aim to investigate these characteristics of the proposed model pipeline through explicit experiments.

### 5.4.1. Subject-Wise Structure of Recognition Performance

Aggregate recognition score can obscure systematic performance differences between subjects. If recognition success is largely determined by a subset of subjects, average performance alone does not capture the model's behaviour. This motivates a subject-wise analysis of recognition outcomes.

As could be observed in Table A.1 mentioned in the primary result, the recognition accuracy is not evenly distributed across all subjects. While some perform well, some are not correctly

recognized even once. This leads to the following hypotheses to concretely quantify this observation:

**H9** Subject-wise recognition performance shows limited but non-random consistency across repeated split observations.

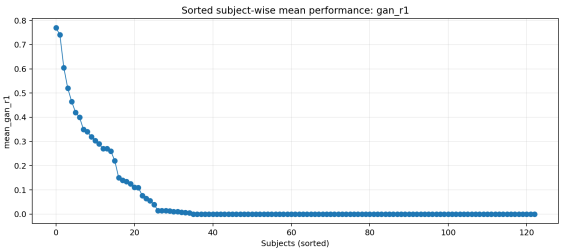
**H10** The observed between-subject heterogeneity in recognition accuracy is explained by the available physiological subject attributes.

#### 5.4.1.1. Experiment

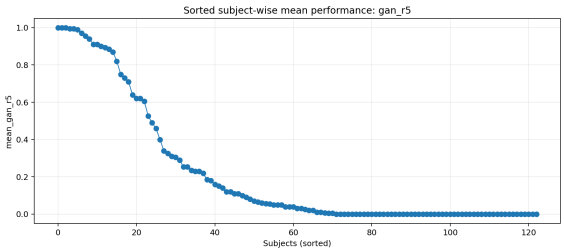
The subject-wise analysis is carried out across all splits. The recognition results for each of the five short-run models are collected. These represent repeated observations for some subjects, although the repetition structure remains sparse. The data contains 190 observations from 123 distinct subjects, with a mean of ~ 1.5 repeated observations per subject. A second part of the analysis is carried out only on the subjects with a non-zero recognition rate. This leads to 45 observations from 35 subjects for Rank-1 accuracy and 103 observations from 71 subjects for Rank-5 accuracy.

First, subjects that are present in multiple splits are investigated. Their mean and standard deviation are computed to assess the stability of recognition accuracies for a given subject across splits. Additionally, the conditional probability of achieving an accuracy greater than 0, given that this subject has an accuracy greater than 0 in another split, is calculated and analyzed.

The subsequent analysis proceeds in two stages. In the first stage, recognition outcomes are reduced to binary hit indicators. A value of 1 indicates that a successful match occurred for a given subject in a given split, while 0 indicates complete failure. This stage addresses the question of basic recognizability. In the second stage, the analysis is restricted to observations with strictly positive metric values. This second stage addresses the question of how strong the recognition result is once a positive case is already present.




**(a)** Subject-wise mean Rank-1 accuracy. Subjects are sorted in descending order of their recognition accuracy across available splits.




**(b)** Subject-wise mean Rank-5 accuracy. Subjects are sorted in descending order of their mean score across available splits.

**Figure 5.1.:** Subject-wise mean recognition accuracy across available splits. The plots show that performance is not evenly distributed across subjects but is concentrated in a subset.

First, mixed-effects and generalized estimating equations (GEE) based models are estimated for the hit-based analysis. For the binary outcome, binomial mixed-effects models are fitted with split-specific fixed effects and a subject-level random intercept. The fixed effects estimate the systematic influence shared across the dataset, which allows each split to have its own baseline recognizability. The random intercept, on the other hand, captures additional subject-specific variation by allowing each subject to deviate from the overall average hit probability. This quantifies whether some subjects are systematically easier or harder to recognize than others. Additionally, GEE models are estimated as a robustness-oriented analysis over all subjects. In contrast to mixed-effect models, GEEs do not explicitly model subject-specific random effects. Instead, they estimate average associations while accounting for repeated observations within the same subject through robust standard errors. This allows the analysis of average attribute effects across all subjects.

Second, the positive-only analysis is modelled using methods appropriate for bounded outcomes in the open interval $(0, 1)$. Zero-valued observations are removed. Because the resulting outcomes are bounded fractions, the primary model is a fractional-logit GEE, which preserves the population-averaged repeated-measures framework used in the hit-based analysis while accommodating fractional responses.

Across both stages, available subject attributes were standardized and entered as explanatory variables. Those include age, height, weight, and shoe size

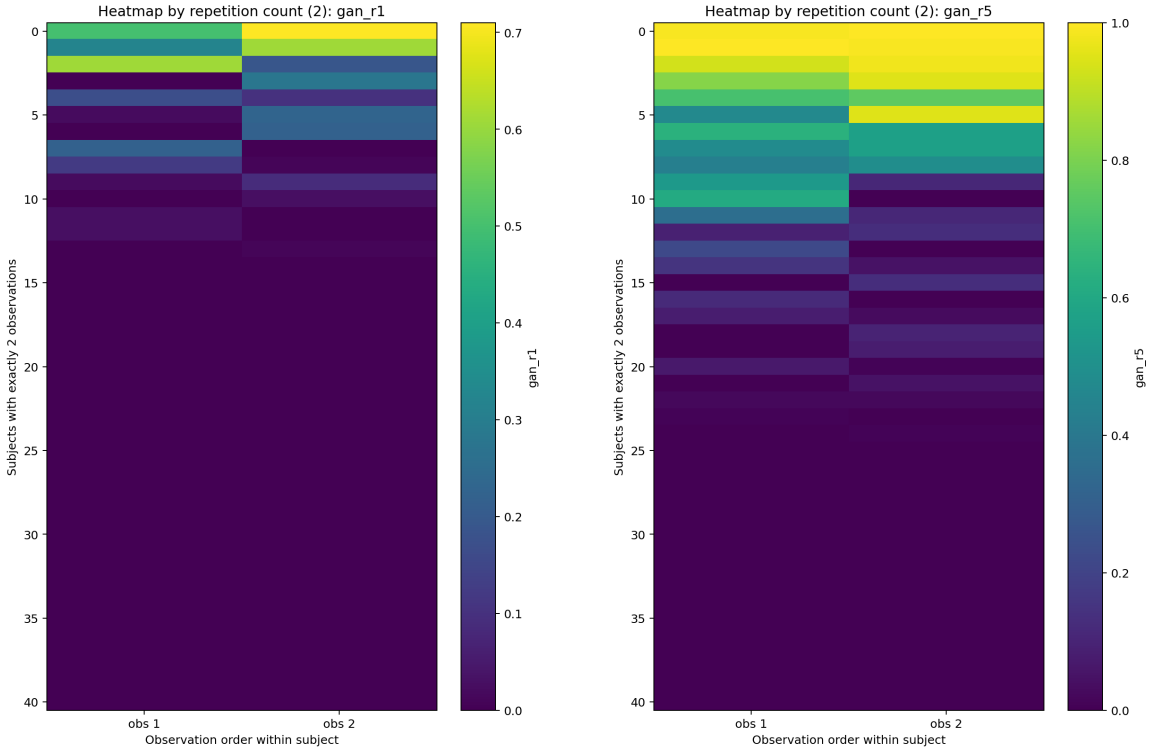
#### 5.4.1.2. Results

The descriptive analysis shows that recognition performance is not evenly distributed across subjects. Instead, successful recognition focuses on a subset of subjects, whereas others achieve only a few or no successful matches across the available splits. This pattern is visible in the sorted subject-wise mean plots visualized in Figure 5.1. Additionally, the heatmaps for subjects with exactly two observations (Figure 5.2) further suggest that the sparse repetition structure is not merely random scatter, especially in Rank-5 accuracies.

The evaluation of the accuracy for subjects with multiple observations is shown in Table 5.6. It is differentiated between 3 different settings. First, all repeated subjects. Second, subjects with at least one occurrence of a non-zero accuracy. Finally, only subjects with no zero

**Table 5.6.:** Subject-level accuracy summary across different subject subsets.

| Metric | Subject subset | $n$ | Mean | Avg Subject Std. dev. |
|---|---|---|---|---|
| Rank-1 | all repeated subjects | 53 | 0.0622 | 0.0342 |
| Rank-5 | all repeated subjects | 53 | 0.2298 | 0.0641 |
| Rank-1 | at least one nonzero | 21 | 0.1569 | 0.1007 |
| Rank-5 | at least one nonzero | 37 | 0.3290 | 0.0970 |
| Rank-1 | no zero observations | 8 | 0.2378 | 0.1321 |
| Rank-5 | no zero observations | 21 | 0.3940 | 0.0944 |




**(a)** Rank-1 **(b)** Rank-5

**Figure 5.2.:** Subject-wise heatmaps for subjects with exactly two observations. Rows correspond to subjects sorted by mean performance, and columns indicate within-subject observation order rather than concrete split identity.

**Table 5.7.:** Recurrence of subject-level success across repeated observations. Pairwise probability denotes $P(\text{success in another observation} \mid \text{success in one observation})$, and baseline probability denotes $P(\text{ success in another observation })$.

| Metric | Pairwise prob. | Baseline prob. |
|---|---|---|
| Rank-1 | 0.595 | 0.223 |
| Rank-5 | 0.843 | 0.614 |

accuracy at all. Additionally, the number of subjects in each category is provided. A clear increase in mean accuracy is observed across the different subsets. For Rank-1, the accuracy increases from 0.0622 to 0.2378, and for Rank-5, it increases from 0.2298 to 0.3940. At the same time, increases are observed in the standard deviations, except for the step from at least one nonzero accuracy to no zero accuracy for Rank-5. On the other hand, in Table 5.7, the results for the probability of success under these repeated observations are shown. For both accuracies, a clear increase is observed from the baseline probability to the pairwise probability. For Rank-1 this increase is from 0.223 to 0.595 while for Rank-5 it is from 0.641 to 0.843

**Table 5.8.:** Hit-based mixed-model summary with subject-level random-intercept heterogeneity.

| Outcome | $SD_{null}$ | $SD_{attr}$ | $\Delta SD$ (%) |
|---|---|---|---|
| Rank-1 hit | 2.226 | 2.117 | 4.9 |
| Rank-5 hit | 3.147 | 3.078 | 2.2 |

SD = standard deviation of the subject-level random intercept. Attr = model with age, height, weight, and shoe size. ΔSD denotes the relative reduction in subject-level standard deviation from the null model to the attribute-extended model.

**Table 5.9.:** Population-averaged coefficient estimates from the hit-based GEE models.

| Outcome | Attribute | Estimate | 95% CI | $p$ |
|---|---|---|---|---|
| Rank-1 hit | age | -0.195 | [-0.589, 0.200] | 0.333 |
| Rank-1 hit | height | -0.234 | [-0.941, 0.472] | 0.516 |
| Rank-1 hit | weight | 0.972 | [0.308, 1.636] | 0.004 |
| Rank-1 hit | shoesize | -0.039 | [-0.747, 0.669] | 0.914 |
| Rank-5 hit | age | -0.024 | [-0.392, 0.343] | 0.896 |
| Rank-5 hit | height | -0.209 | [-0.919, 0.501] | 0.564 |
| Rank-5 hit | weight | 0.442 | [-0.161, 1.046] | 0.151 |
| Rank-5 hit | shoesize | 0.266 | [-0.455, 0.987] | 0.469 |

The hit-based models confirm the pattern of subject-heterogeneity quantitatively. In the hit-based GEE models, both outcomes were estimated on 189 observations from 122 subjects, which provides a common inferential basis for the first-stage analysis (one observation was removed as no weight or height was available). The corresponding mixed-effects models between-subject heterogeneity as seen in Table 5.8: in the null model for Rank-1 accuracy, the estimated standard deviation of the subject-level random effect is 2.226 and for Rank-5 accuracy 3.147. After adding age, height, weight, and shoe size, these estimates decrease slightly to 2.117 for Rank-1 and 3.078 for Rank-5. In the multivariable GEE analysis, standardized body weight is the only robust attribute associated with recognizability shown in Table 5.9. For Rank-1, standardized weight has a coefficient of 0.9717 with a $p$-value of 0.004 and a 95% confidence interval from 0.308 to 1.636. For Rank-5, no attribute showed a comparably robust effect.

The positive-only analysis shows that the subject-wise structure remains visible even after excluding complete failures. Therefore, the subject-wise differences are not limited to the distinction between recognition and non-recognition, but also persist among successful cases. However, the strength of the evidence differs substantially between Rank-1 and Rank-5.

For positive Rank-1 accuracies, the evidence remains weak and unstable. The subset comprises only 45 observations from 35. The multivariable fractional-logit GEE fails numerically and returns no interpretable parameter estimates. Therefore, no reliable inference can be drawn.

**Table 5.10.:** Population-averaged coefficient estimates from positive-only fractional-logit GEE models.

| Outcome | Attribute | Estimate | 95% CI | $p$ |
|---|---|---|---|---|
| Rank-5 | age | -0.014 | [-0.352, 0.324] | 0.936 |
| Rank-5 | height | -0.182 | [-0.997, 0.633] | 0.662 |
| Rank-5 | weight | 0.727 | [0.099, 1.355] | 0.023 |
| Rank-5 | shoesize | -0.419 | [-1.198, 0.359] | 0.291 |

Positive Rank-1 fractional-logit GEE did not converge and therefore yielded no interpretable multivariable coefficient estimates.

For positive Rank-5 accuracies, the GEE model returns parameter estimates, shown in Table 5.10. The positive-only Rank-5 subset consists of 103 observations from 71 subjects. Standardized weight is the only significant positive predictor, with a coefficient of 0.727, a $p$-value of 0.023, and a 95% confidence interval from 0.099 to 0.1355. The remaining attributes do not show robust effects in either model.

#### 5.4.1.3. Findings

To summarize the result, the combined subject-wise analysis shows that recognition performance is structured at two levels. First, subjects differ in their basic recognition ability. Some subjects are repeatedly recognized, while others remain unrecognized across different data splits. This can be seen visually in the plots but is also quantified in the estimated standard deviation of the subject-level random effect. Second, these differences persist even after restricting the analysis to successful cases, indicating that subject-wise structure is not limited to complete recognition failure but also affects the strength of positive recognition accuracy. The visualizations and the inferential models consistently support this conclusion.

Additionally, the investigation of the subjects' means and standard deviations across the splits indicates that results for subjects who are repeatedly present across splits are heavily influenced by subjects with zero accuracy in one or more observations. This is reflected in the low mean accuracies and mean standard deviations. Once subjects with nonzero accuracies are filtered out, the mean accuracy increases, but the observed within-subject variability also increases. This suggests that the accuracy for subjects is less stable across splits than would be expected from a subset of all subjects with multiple observations. However, when considering only successful cases, the conditional probability of success in another observation under the condition is higher than the overall success probability. These results suggest some subject-level consistency across splits, since subjects that are successful once are more likely to be successful again than expected from the overall success rate. This therefore supports H9.

At the same time, the investigated subject attributes explain only a limited portion of the structure of subject-level heterogeneity. In the hit-based model, the inclusion of the attributes reduced the estimated subject-level heterogeneity by only 0.109 for Rank-1 and

0.069 for Rank-5. In the positive-only analysis, the same general pattern remained. This indicates that a substantial portion of the subject-wise variation cannot be explained by the investigated attributes, and therefore H10 cannot be confidently supported.

Across both stages, body weight is the only attribute that exhibits a robust, consistent association with subject-wise performance. In the hit-based analysis, weight is positively associated with the probability of a Rank-1 hit, but not with the probability of a Rank-5 hit. However, in the positive-only analysis, no robust effect was observed for positive Rank-1 performance, whereas weight was a consistent positive predictor of positive Rank-5 performance in the GEE. Therefore, the most defensible overall conclusion is that subject-wise recognition performance is structured, but can only be partially explained by the investigated subject attributes.

Finally, it needs to be mentioned that these findings should be interpreted with appropriate caution. The repeated-measures structure is relatively thin, especially in the positive Rank-1 subset, and most subjects appear only once or twice. The present analysis is well-suited to establish the presence of subject-specific heterogeneity, but less suited to support strong claims about fine-grained within-subject stability. Because height, weight, and shoe size are moderately to strongly correlated, individual coefficient estimates should be interpreted as conditional associations rather than isolated physiological effects.

### 5.4.2. Image-Level Similarity vs. Recognition

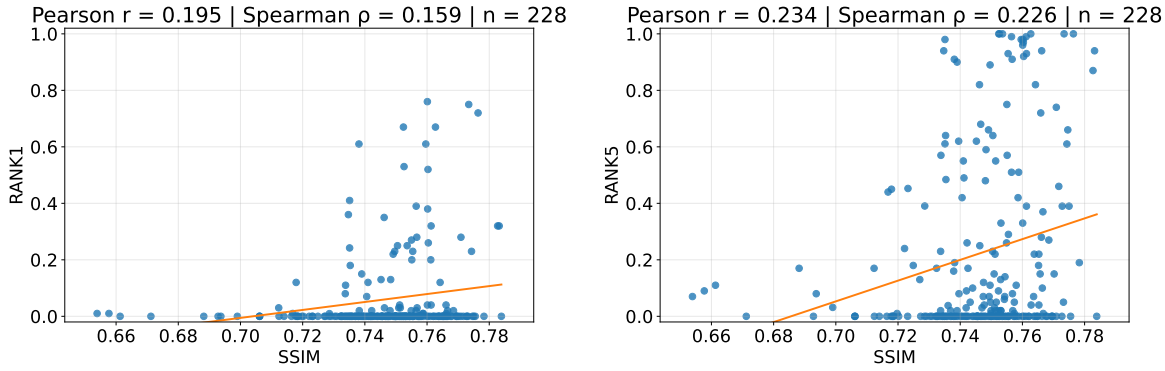
The training objective and the evaluation protocol assess the model from different viewpoints. While image-level metrics are a substantial part of the objective function, the downstream recognition model operates in a learned embedding space optimized for identity discrimination. Therefore, it is unclear whether improvements in image-level similarity translate into better recognition performance, or vice versa.

**H11** Image-level reconstruction quality and subject-level recognition performance are only weakly associated. Improvements in one do not reliably translate into improvements in the other.

#### 5.4.2.1. Experiment

To analyse the relationship between reconstruction quality and downstream identification, the models from the five short training runs and the long training run are evaluated.

For each evaluated model, image-level reconstruction quality is measured on the test set using the Structural Similarity Index Measure (SSIM), as it is not directly captured in a loss term that is optimized. Downstream recognition performance, on the other hand, is measured using the same gallery-probe identification protocol as in the primary experiment. As single averaged values for each model would provide few data points for analysis, the subject-wise results are used.




**(a)** SSIM versus subject-level Rank-1 accuracy.

**(b)** SSIM versus subject-level Rank-5 accuracy.

**Figure 5.3.:** Subject-level relationship between silhouette similarity and downstream recognition performance. Each point represents one subject-level observation.

The analysis examines whether changes in a model's image-level scores also affect Rank-1 or Rank-5 accuracies. The relationship is quantified using Spearman correlation, a rank-based statistical association measure. Additionally, it differs between one setting where all observations are used and one where only positive observations are used.

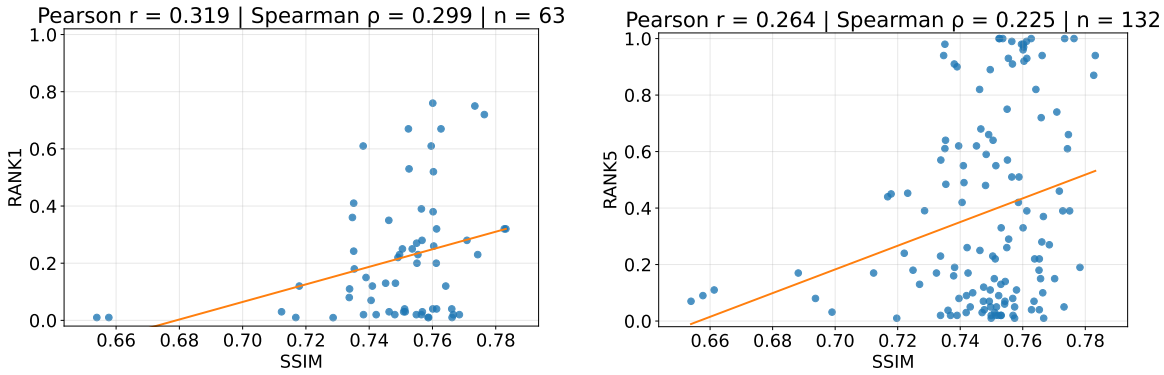
#### 5.4.2.2. Results

The subject-wise analysis yields 288 observations. The scatter plots for Rank-1 and Rank-5 accuracy against SSIM can be seen in Figure 5.3. For Rank-1 accuracy, a Pearson correlation coefficient of $r = 0.195$ and a Spearman rank correlation of $\rho = 0.159$ exists. For Rank-5 accuracy, Pearson correlation coefficient $r = 0.234$ and Spearman rank correlation $\rho = 0.226$ are obtained.

The subject-wise analysis, restricted to positive observations, yields 63 observations for Rank-1 and 132 for Rank-5. The scatter plots for Rank-1 and Rank-5 accuracy against SSIM can be seen in Figure 5.4. For Rank-1 accuracy, a Pearson correlation coefficient of $r = 0.319$ and a Spearman rank correlation of $\rho = 0.299$ exists. For Rank-5 accuracy, Pearson correlation coefficient $r = 0.264$ and Spearman rank correlation $\rho = 0.225$ are obtained.

#### 5.4.2.3. Findings

The subject-based analysis suggests that SSIM captures only a limited portion of the information relevant to downstream recognition. For both Rank-1 and Rank-5 accuracies, a weak positive association with SSIM is observed. This indicates that subjects with a higher structural similarity between generated and reference silhouettes tend, on average, to achieve slightly better recognition performance. However, this effect is only small.




**(a)** SSIM versus subject-level Rank-1 accuracy.

**(b)** SSIM versus subject-level Rank-5 accuracy.

**Figure 5.4.:** Subject-level relationship between silhouette similarity and downstream recognition performance of only positive accuracies. Each point represents one subject-level observation.

The relationship for Rank-5 is slightly stronger than for Rank-1. This suggests that improvements in structural similarity contribute more to broader retrieval than to exact Rank-1 identification.

When considering only positive accuracies, the Rank-1 case is clearly dominated by zero-accuracy cases, as the association with SSIM increases once these cases are removed. This is not the case for Rank-5, as here it remains nearly unchanged. This suggests that SSIM is somewhat more related to the strength of recognition once Rank-1 identification is achieved.

Although small positive associations are observed, SSIM cannot be interpreted as a reliable proxy for the underlying recognition accuracy. The weak correlation may indicate that structurally similar silhouette reconstructions do not preserve identity-relevant gait characteristics. A possible explanation is that SSIM primarily captures global structural similarity, whereas important fine-grained gait information has only a limited influence on it.

All together, these findings support the conclusion that SSIM has at most a limited influence on the quality of generated silhouette sequences for downstream gait recognition. This supports H11, because although a small relationship exists, it cannot be used to reliably translate one improvement into the other.

### 5.4.3. Temporal Motion Features

Visual inspection of the generated sequences suggests that the movement directions are preserved well. Because the other information in the point cloud primarily provides static information, the motion direction may depend on motion-related radar features and on the timestamps associated with each point.

**Table 5.11.:** Aggregated frame-wise walking-direction assessment across all frames. Correct, reverse, and unclear are reported relative to all frames.

| Configuration | Corr. (%) | Rev. (%) | Unclear (%) |
|---|---|---|---|
| normal | 98.33 | 1.67 | 0.00 |
| vel-rem | 85.28 | 7.50 | 7.22 |
| time-rem | 71.11 | 13.06 | 15.83 |
| both-rem | 49.44 | 22.50 | 28.06 |
| vel-inv. | 66.94 | 24.72 | 8.33 |
| time-inv. | 31.39 | 57.50 | 11.11 |
| both-inv | 1.11 | 88.06 | 10.83 |

**H12** Movement direction is not recovered from the static pose alone. Instead, the generator relies primarily on velocity and timestamp features in the radar input to reconstruct directionally consistent motion.

#### 5.4.3.1. Experiment

A controlled feature ablation is performed by removing velocity and timestamp information from the radar input, while keeping all other pipeline components unchanged during inference. This is achieved by setting the velocity or timestamp input for each point to 0. Additional variants either remove only velocity or only timestamp information, or flip the sign of the input. These variants will be named as vel-rem, time-rem, both-rem, vel-inv, time-inv, both-inv to keep later sections concise.

All variants are evaluated under the same first data split. Generated sequences are analyzed with respect to their movement direction in the depth reference frame, frame-by-frame. Directional consistency is then categorized as correct, incorrect, or not clearly definable. This is evaluated only on a random sample of 10 recordings (yielding 723 frames), as it is done by hand because a concrete labeling is not available, and unclear results may be difficult to categorize otherwise.

#### 5.4.3.2. Results

In Table 5.11, the relative results of all experiments can be seen. For each configuration, the percentages of correct, reversed, and unclear walking directions are reported relative to the ground-truth frames. Additionally, percentages for correct and reversed walking directions are reported across all clearly categorizable frames. The highest percentage of correct directions is observed in the normal configuration, at 98.33%, whereas the lowest is in the both-inv configuration, at 1.11%. For reversed walking directions, this pattern is switched. The normal configuration has the lowest percentage, 1.67%, whereas both-inv achieves 88.06%. The highest percentage of unclear walking directions is observed in both-rem, at

28.06%, while the lowest is in the normal configuration, at 0.00%. The cleaned percentages do not change the ranking of the configurations for correct and reversed walking direction.

#### 5.4.3.3. Findings

The presented results indicate that the model's default configuration has learned to correctly infer walking direction from the point cloud input, as reversed walking directions occur in only 1.67% of frames and no ambiguous walking directions are observed. This suggests that the generated silhouettes largely preserve the correct directional movement.

However, manipulation of velocity or timestamp values significantly affects the walking direction. Removing either velocity or timestamp information results in a consistent decrease in correct walking directions and an increase in reversed and unclear directions. This effect is more pronounced for time-rem than for vel-rem, indicating that timestamp information is more important for determining the correct walking direction than velocity. Velocity still contributes to the walking direction; however, it is clearly less decisive.

Analyzing the inverse configurations yields an even clearer picture. For both vel-inv and time-inv, the percentage of reversed walking direction substantially increases, whereas the change in unclear walking direction remains comparatively limited. Therefore, inversion does not primarily lead to unclear walking directions, but rather to a shift in the opposite direction. As this is also stronger for timestamp inversion, this again suggests that the model uses timestamp information as the primary cue for direction derivation.

The strongest insight is delivered by inverting both simultaneously. Here, only 1.11% of frames depict the correct walking directions, while the percentage of unclear frames stays similar to the other inversion configurations. Especially when compared to the unclear percentage, nearly all categorizable frames showed a reversed direction. As the unclear percentage remains comparatively stable, the joint inversion does not degrade the walking direction but mainly flips it.

The remaining unclear frames should not be interpreted directly as evidence against the directional reversal. A possible explanation is that the combination of a reversed timestamp and velocity, with no change in the other attributes, introduces uncertainty in the model, as the timestamp and velocity values no longer match the expected inputs on which it was trained. However, importantly, this uncertainty remains limited.

Another effect is observed when both the timestamp and velocity information are removed. In this case, the percentage uncertainty reaches its maximum of 28.06%, substantially higher than the 10.83% of the both-inv configuration. This suggests that the absence of both values does not directly enforce a reversal of the walking direction but rather reduces the model's ability to produce coherent motion. Therefore, reversal primarily changes the derived direction, whereas removal disrupts the model's ability to produce interpretable movement direction.

# 6. Discussion

This section discusses the relevant findings and their implications. Sections 6.1 summarize the main findings and bridge the gap to concrete research questions. Section 6.2 discusses concrete implications for further research. Section 6.3 analyzes the implications those findings have in practice. Section 6.4 describes the existing limitations of this thesis, while Section 6.5 covers possible future directions and works.

## 6.1. Principal Findings

This section discusses the results regarding the central research question of this thesis. The findings show that, first, the proposed radar-to-depth pipeline indeed enables identification above the chance level. Therefore, this supports the fundamental hypothesis that radar point-cloud-based observations can be translated into a format at least partially compatible with an existing depth-based gait-recognition model.

Second, the gap to accuracy on real-depth data, however, remains substantial. It demonstrates no practical, efficient substitute approach for radar-based recognition, but rather a limited but technically feasible workaround. This interpretation is especially relevant from a data privacy perspective. For the adversary model, no strong attack is needed. Already, the confirmation that an identity-relevant gait characteristic can be partially reconstructed from radar data and used for downstream recognition demonstrates that the switch from camera-based to radar-based sensors does not reliably eliminate the possibility of identification.

Third, the results indicate that the bottleneck lies in our generative translation, not in the downstream recognition model. The nearly perfect performance of the depth-based gait recognition model on real depth silhouettes shows that the evaluation protocol is functional. The performance reduction for synthetic probes can therefore be interpreted as information loss during the translation from radar point clouds to depth silhouettes.

Fourth, the weak correlation between SSIM and recognition accuracy indicates that image-related reconstruction and biometric identification are only loosely related. In the present case, it is not decisive whether synthetic images are visually plausible, but rather whether gait-relevant, discriminative structures are preserved during translation.

Finally, the ablation studies regarding velocity and timestamp attributes suggest that the model does not generate only static silhouettes. In contrast, temporal and velocity information are particularly used to reconstruct movement directions. This is relevant for gait recognition as gait is a dynamic pattern that depends on temporal consistency.

In relation to the state of research, this thesis is positioned between two established lines of thought. On the one hand, it differs from two-stream embedding alignment architectures that assume identity supervision. On the other hand, it goes beyond mere feasibility of radar-based depth generation for pose estimation, as not the reconstruction of human pose but usability for downstream recognition is in focus. Methodologically, the approach is therefore closest to biometric translation strategies such as cross-spectral face recognition: if an input modality does not fit the recognition model, generating the native input modality for the existing model is an alternative to training a new model.

### 6.1.1. Discussion with Respect to Research Questions

Taken together, these findings allow the research questions of this thesis to be answered more precisely.

The results of this thesis show that existing GAN principles can be adapted to radar-to-depth translation. This approach is implemented using a PointNet-based encoder for sparse radar point clouds, masking-aware pooling for padded batched inputs, an enriched global latent representation, a FiLM-conditioned convolutional decoder, and a projection-based PatchGAN discriminator. Additionally, the objective function extends standard adversarial image reconstruction by including an overlap-based and gait-embedding-based loss term. However, the ablation result indicates that not all adopted GAN components contribute equally. While FiLM conditioning appears to be particularly important, adversarial training itself provides no clear benefit under the present setup.

The thesis demonstrates that privacy implications of cross-sensor gait recognition can be evaluated empirically by translating legal or organizational privacy concerns into explicit adversary models and measurable identification protocols. Concretely, the present work models a setting in which radar is permitted while camera-based sensing is restricted, but a pre-existing depth-based recognition model and a paired radar-depth dataset are available. The privacy risk is operationalized through a gallery-probe identification experiment using real depth galleries and from radar-translated synthetic depth probes. This leads to privacy implications not being discussed abstractly, but tested through measurable downstream identification performance under subject-disjoint evaluation.

The empirical results show that machine learning can be used for identity-preserving domain translation to a limited but non-trivial extent. Across all evaluation models, downstream reconstruction is enabled above chance level. This indicates that translation preserves part of identity-relevant gait information. However, a large gap to real-depth references indicates that preservation remains incomplete. This supports feasibility but not high fidelity for identity preservation. It therefore provides evidence for partial rather than identity-preserving translation.

The success of cross-sensor gait recognition in the present setting is influenced by multiple factors. First, the temporal context. Smaller and larger aggregations reduce performance relative to the chosen baseline. Second, temporal data, in the form of normalized timestamps, are less significant for downstream recognition but highly relevant for reconstructing motion direction. Additionally, the generator's architectural design is critical. The FiLM-based decoder conditioning shows a clear positive contribution. However, adversarial loss does not seem to be essential under the present setting. Image-level similarity only seems weakly related to downstream identification success. This indicates that visually plausible translation is not sufficient for recognition performance. Finally, recognition performance varies substantially across subjects, while the heterogeneity can only be partially explained by the available subject attributes.

## 6.2. Implications for Research

For research purposes, this thesis proposes generative translation as an independent paradigm for cross-sensor gait recognition. Existing work on cross-sensor recognition relies overwhelmingly on explicit identity supervision. This thesis, on the other hand, shows that even paired but unlabeled data in a radar-depth setting can be sufficient for a limited but measurable path to identification. This is particularly significant when identity labels are unavailable or not realistically obtainable.

Furthermore, the findings imply the need for a task-specific objective function. Reconstruction metrics or overlap metrics alone are insufficient for biometric translation, as they do not directly measure identity-relevant information. This thesis, therefore, argues for a stronger connection between the generative model and downstream recognition. This can be achieved either through recognition-based losses or through evaluation that explicitly distinguishes between visual quality and recognition performance.

## 6.3. Implications for Practice

For practice, this thesis shows that sensor-based privacy gaps can be more tractable than previously assumed. If depth cameras are prohibited but radar is allowed, this does not necessarily imply that biometric identification is not possible. As soon as a paired radar-depth dataset and an existing depth recognition model are available, a generative translation can, in principle, reconstruct the prohibited modality at least partially. However, at the same time, those results should not be over-interpreted. The achieved recognition accuracy is far from sufficient for robust surveillance use. The practical relevance, therefore, lies less in an immediately applicable offensive technique than in a warning about a potential risk pathway that has so far been plausible but not explicitly investigated.

This implies that not only collections of image-based data should be regulated, but also subsequent recombinations of these data with different sensor data. Especially important

are paired calibrations and test data, which may seem to serve only technical purposes but can help bridge the gap between permitted and prohibited sensor modalities. Likewise, existing gait recognition models should be considered part of the attack surface. Earlier, legitimately trained depth recognition models could be reused in another sensor setting, although the original sensor modality is no longer permitted.

## 6.4. Limitations

While this thesis has its relevant takeaways, some limitations also warrant mention. First, the evaluation is not based on strictly disjoint datasets. While the adversary model is about an existing trained depth-based gait recognition model, this explicit disconnect cannot be made here. The training and test subjects are disjoint. However, the reconstruction and generative models both train on the same dataset. While the test subjects are unseen by both, the results show a compatibility under controlled dataset coherency, but not strict robustness to an externally trained depth recognition model.

Second, the frame alignment during preprocessing is only approximate. While the radar-depth data are paired at the recording level and share a synchronized start time, the frame alignment is index-based and, due to a slight frame rate mismatch, only approximate. Although this accounts for only a small shift, it remains as an additional noise source in the learning signal, especially for our approach with a frame-wise translation model.

Third, the gait-related loss does not reflect the downstream task perfectly. Although the downstream identification is realized via nearest-neighbor using a similarity or distance measure, the auxiliary gait loss performs a mean-squared error operation between the generated and the ground truth embeddings. This creates a mismatch between the optimization target and the concrete metric used during gait recognition. An exploratory result from a single additional run on the first data split suggests that such a more aligned gait loss is relevant for successful retrieval. The additional run shows an improvement of Rank-1 accuracy from 0.066 to 0.096. However, because the observation is based on only one split, it does not necessarily constitute a confirmed finding and should be interpreted as an indication.

Fourth, a central methodological limitation is the disconnect between frame-wise generation and sequence-based recognition. While the generator optimizes reconstruction on a single image, the downstream reconstruction processes complete gait sequences. Therefore, a structural gap exists between the training and evaluation objectives.

Fifth, hyperparameter tuning and the final evaluation are not strictly performed on subject-disjoint partitions. As the hyperparameter tuning is done on a random data split, it can not be ensured that subjects used for during the tuning process are not in the test set during evaluation. Therefore, a slight optimistic bias can not be excluded.

Finally, the dataset used could be improved in several ways. The limited dataset size is a concern, as it effectively constrained the training of the recognition model on the same

dataset. Additionally, a larger dataset could enable different models, such as sequence-to-sequence, that directly train on pairs of recordings and might need for data. Furthermore, the radar point clouds are very sparse and noisy. This is, in general, an issue with radar point clouds; however, it introduces the problem that the constraint under which the generator creates its output is information-sparse. Especially in regions that are highly influential on the silhouette, such as arms and legs, the point cloud data are relatively sparse.

## 6.5. Future Work

Possible research directions for future work could investigate approaches that remain unanswered or were deliberately not chosen. Those follow directly from the results and limitations of this thesis.

A first direction concerns temporal modeling. The current approach follows a frame-wise generation approach, whereas the downstream recognition relies on full silhouette sequences. Therefore, sequence-level approaches, such as vid2vid [45] could be investigated. As mentioned in Section 4.2, we did not choose this approach because it does not remove the issue of generating a structurally feasible silhouette. However, it can introduce a more temporally consistent sequence of frames, but will also increase the model's complexity. Additionally, more ambitious sequence-level approaches could be investigated: processing the entire point-cloud sequence and generating a complete depth-silhouette sequence rather than individual frames. This would align the generative model more closely with downstream recognition, but would substantially increase complexity and may require more data, as they would have to learn a richer, more difficult mapping. The latter one is already an issue as discussed in Section 6.4.

A second direction concerns the training objective. Future work can investigate the model's inaccuracies, particularly the limitations of the gait-embedding-based loss, as it is currently based on mean squared error. This could systematically examine the influence of choosing a metric more aligned with the actual retrieval metric on the performance of the downstream task. This approach would still satisfy all assumptions of the adversary model. However, it does not address the potential issue of the quality of an embedding of a single frame.

A third direction is to reconsider the underlying generative paradigm. As indicated by the ablation studies in Section 5.3, adopting a completely different generative model paradigm might be worthwhile. The removal of the adversarial loss suggests that including an adversarial component in our approach may even be harmful. Therefore, investigating non-adversarial generative approaches could be insightful.

A fourth direction, to address a limitation of this thesis, would be to explicitly apply an external, independently trained depth recognizer. This can give insights into whether the results of this thesis are data-specific or generalize beyond the current experimental setting to other datasets.

At last, privacy-enhancing technologies need to be mentioned. While privacy-enhancing technologies are an established field, they may not be evaluated with respect to this attack vector. Although our methodology yields only limited success, it demonstrates that gait-relevant information can be translated across sensor domains to a measurable extent. Therefore, evaluating existing privacy-enhancing technologies for such a setting may be suitable.

# 7. Conclusion

This thesis investigates radar-to-depth cross-sensor gait recognition via GANs. More precisely, whether mmWave radar point-cloud sequences can be converted into depth-silhouette sequences while retaining enough discriminative gait characteristics for downstream gait recognition. If this is the case, it would enable bridging the gap from radar to depth sensors and enable gait recognition without the need for an explicitly labeled radar dataset. This would, therefore, argue against the perceived privacy-friendliness of radar sensors.

The results show that a radar-to-depth translation is feasible and enables the downstream depth-based gait recognition to a limited but measurable extent above chance level. However, the performance still shows a substantial gap relative to the depth-based gait recognition model on real depth recordings. Therefore, while some identity-relevant gait characteristics are translated, not enough are preserved to reliably perform gait recognition.

The core empirical findings of this thesis indicate that the proposed method works above chance across different data splits, while the gap to recognition on real depth data remains very large. The FiLM conditioning in the decoder proves to be an important component for the performance of the downstream recognition. The adversarial loss, on the other hand, shows no clear benefit and suggests that non-adversarial approaches are a promising direction for future work. Image similarity between the synthetic and ground-truth silhouettes shows only a weak relationship with the resulting accuracy, suggesting that greater emphasis on objectives that do not prioritize image similarity might be appropriate. Additionally, examining subject-wise performance shows that accuracy is highly heterogeneous across subjects.

The main implication of this thesis regarding privacy concerns is that radar may be used to leak identity-relevant gait information across sensor boundaries. Replacing cameras with radar does not automatically remove biometric privacy risk if no identity-labeled radar dataset is present. Already, the availability of a paired radar-depth dataset and an existing depth-based gait recognition model can bridge this gap. Using a GAN-based translation model, the observed radar data can be translated into depth data for downstream recognition. This indicates the need to evaluate privacy claims for radar sensors under such translation-based cross-sensor attack vectors.

To summarize these points, translating radar data into depth data for use as input to a depth-based gait recognition model is technically feasible. However, in practice, it shows substantial limitations, especially compared to real depth sequence recognition. Still, the insight that this translation is feasible provides privacy-relevant implications. Therefore, this approach and its potential privacy implications deserve further research.

# Bibliography


[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 2623–2631. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701.

[2] David Anghelone, Cunjian Chen, Arun Ross, and Antitza Dantcheva. *Beyond the Visible: A Survey on Cross-spectral Face Recognition.* arXiv:2201.04435 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2201.04435.

[3] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. "Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (Sept. 2022), 102:1–102:25. DOI: 10.1145/3550325.

[4] Cunjian Chen and Arun Ross. "Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network". In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. May 2019, pp. 1–8. DOI: 10.1109/FG.2019.8756527.

[5] Hao Chen et al. "Wi-GR: Wi-Fi-Based Gait Recognition Using Multi-Part Velocity Profile". In: *IEEE Transactions on Mobile Computing* 24.11 (Nov. 2025), pp. 11957–11971. ISSN: 1558-0660. DOI: 10.1109/TMC.2025.3581549.

[6] Yuwei Cheng and Yimin Liu. "Person Reidentification Based on Automotive Radar Point Clouds". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–13. ISSN: 1558-0644. DOI: 10.1109/TGRS.2021.3073664.

[7] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. "Cross-Modality Person Re-Identification with Generative Adversarial Training". en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 677–683. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/94.

[8] Lang Deng, Jifang Pei, Yuansen Song, Weibo Huo, Yin Zhang, and Yulin Huang. "KiRV: Robust Human Identification via Multimodal Learning Based on Kinetic Gait Features of Radar and Vision". In: *IEEE Internet of Things Journal* 12.12 (June 2025), pp. 22224–22242. ISSN: 2327-4662. DOI: 10.1109/JIOT.2025.3550532.

[9] Yunze Deng, Haijun Xiong, and Bin Feng. "Licaf: Lidar-Camera Asymmetric Fusion For Gait Recognition". In: *2024 IEEE International Conference on Image Processing (ICIP)*. Oct. 2024, pp. 2424–2430. DOI: 10.1109/ICIP51287.2024.10647273.

[10] Xing Di, He Zhang, and Vishal M. Patel. "Polarimetric Thermal to Visible Face Verification via Attribute Preserved Synthesis". In: *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. Oct. 2018, pp. 1–10. DOI: 10.1109/BTAS.2018.8698554.

[11] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 0012-9658. DOI: 10.2307/1932409.

[12] Tom Eelbode et al. "Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index". en. In: *IEEE Transactions on Medical Imaging* 39.11 (Nov. 2020). arXiv:2010.13499 [eess], pp. 3679–3690. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2020.3002417.

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.

[14] Junqiao Fan, Haocong Rao, and Xuehe Wang. "From Radar to Depth: Multi - Modality Co- Learning for High-Resolution Human Pose Estimation". In: *2024 IEEE International Conference on Smart Internet of Things (SmartIoT)*. Nov. 2024, pp. 248–253. DOI: 10.1109/SmartIoT62235.2024.00044.

[15] Ian J. Goodfellow et al. *Generative Adversarial Networks*. arXiv:1406.2661 [stat]. June 2014. DOI: 10.48550/arXiv.1406.2661.

[16] Mehmet Günel, Erkut Erdem, and Aykut Erdem. *Language Guided Fashion Image Manipulation with Feature-wise Transformations*. arXiv:1808.04000 [cs]. Aug. 2018. DOI: 10.48550/arXiv.1808.04000.

[17] Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. *Camera-LiDAR Cross-modality Gait Recognition*. arXiv:2407.02038 [cs]. July 2024. DOI: 10.48550/arXiv.2407.02038.

[18] Xun Huang and Serge Belongie. "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization". en. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 1510–1519. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.167.

[19] Seyed Mehdi Iranmanesh and Nasser M. Nasrabadi. "Attribute-Guided Deep Polarimetric Thermal-to-visible Face Recognition". In: *2019 International Conference on Biometrics (ICB)*. June 2019, pp. 1–8. DOI: 10.1109/ICB45273.2019.8987416.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks*. arXiv:1611.07004 [cs]. Nov. 2018. DOI: 10.48550/arXiv.1611.07004.

[21] A.K. Jain, A. Ross, and S. Prabhakar. "An Introduction to Biometric Recognition". en. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14.1 (Jan. 2004), pp. 4–20. ISSN: 1051-8215. DOI: 10.1109/TCSVT.2003.818349.

[22] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* (Dec. 2014).

[23] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. *Least Squares Generative Adversarial Networks*. arXiv:1611.04076 [cs]. Apr. 2017. DOI: 10.48550/arXiv.1611.04076.

[24] Stefan Milz, Martin Simon, Kai Fischer, Maximillian Pöpperl, and Horst-Michael Gross. "Points2Pix: 3D Point-Cloud to Image Translation Using Conditional GANs". en. In: *Pattern Recognition*. Ed. by Gernot A. Fink, Simone Frintrop, and Xiaoyi Jiang. Cham: Springer International Publishing, 2019, pp. 387–400. ISBN: 978-3-030-33676-9. DOI: 10.1007/978-3-030-33676-9_27.

[25] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. arXiv:1411.1784 [cs]. Nov. 2014. DOI: 10.48550/arXiv.1411.1784.

[26] Takeru Miyato and Masanori Koyama. *cGANs with Projection Discriminator*. arXiv:1802.05637 [cs]. Aug. 2018. DOI: 10.48550/arXiv.1802.05637.

[27] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. *FiLM: Visual Reasoning with a General Conditioning Layer*. arXiv:1709.07871 [cs]. Dec. 2017. DOI: 10.48550/arXiv.1709.07871.

[28] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. arXiv:1612.00593 [cs]. Apr. 2017. DOI: 10.48550/arXiv.1612.00593.

[29] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. *PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space*. arXiv:1706.02413 [cs]. June 2017. DOI: 10.48550/arXiv.1706.02413.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv:1505.04597 [cs]. May 2015. DOI: 10.48550/arXiv.1505.04597.

[31] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. *Tversky loss function for image segmentation using 3D fully convolutional deep networks*. en. arXiv:1706.05721 [cs]. June 2017. DOI: 10.48550/arXiv.1706.05721.

[32] Claudio Filipi Gonçalves dos Santos et al. "Gait Recognition Based on Deep Learning: A Survey". In: *ACM Computing Surveys* 55.2 (Feb. 2023). arXiv:2201.03323 [cs], pp. 1–34. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3490235.

[33] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. *LidarGait: Benchmarking 3D Gait Recognition with Point Clouds*. arXiv:2211.10598 [cs]. Mar. 2023. DOI: 10.48550/arXiv.2211.10598.

[34] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q. Huang, and Liang Wang. "A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 7.2 (Apr. 2025), pp. 270–292. ISSN: 2637-6407. DOI: 10.1109/TBIOM.2024.3486345.

[35] Yu Shi et al. "Robust Gait Recognition Based on Deep CNNs With Camera and Radar Sensor Fusion". In: *IEEE Internet of Things Journal* 10.12 (June 2023), pp. 10817–10832. ISSN: 2327-4662. DOI: 10.1109/JIOT.2023.3242417.

[36] Zengyu Song et al. "IMFi: IMU-WiFi based Cross-modal Gait Recognition System with Hot-Deployment". In: *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*. Dec. 2021, pp. 279–286. DOI: 10.1109/MSN53354.2021.00052.

[37] Yash Soni, Malhaar Goswami, Nishit Prabhakar Shetty, and Dhiraj. "Millimeter-wave radar for intelligent sensing: A comprehensive review of techniques, applications, and challenges". In: *Computers and Electrical Engineering* 128 (Dec. 2025), p. 110696. ISSN: 0045-7906. DOI: 10.1016/j.compeleceng.2025.110696.

[38] Thorvald Julius Sørensen. "A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: *Biologiske Skrifter* 5 (1948), pp. 1–34.

[39] Tino Stöckel, Robert Jacksteit, Martin Behrens, Ralf Skripitz, Rainer Bader, and Anett Mau-Moeller. "The mental representation of the human gait in young and older adults". eng. In: *Frontiers in Psychology* 6 (2015), p. 943. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00943.

[40] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations". In: vol. 10553. arXiv:1707.03237 [cs]. 2017, pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28.

[41] Julian Todt, Felix Morsbach, Phillip Dissert, and Thorsten Strufe. *MultiGait: A Multi-Sensor Multi-Session Multi-Perspective Gait Dataset and Benchmark*. Technical Report. 2025.

[42] Amos Tversky. "Features of similarity". In: *Psychological Review* 84.4 (1977), pp. 327–352. ISSN: 1939-1471. DOI: 10.1037/0033-295X.84.4.327.

[43] Changsheng Wan, Li Wang, and Vir V. Phoha. "A Survey on Gait Recognition". In: *ACM Comput. Surv.* 51.5 (Aug. 2018), 89:1–89:35. ISSN: 0360-0300. DOI: 10.1145/3230633.

[44] Rui Wang, Chuanfu Shen, Manuel J. Marin-Jimenez, George Q. Huang, and Shiqi Yu. "Cross-Modality Gait Recognition: Bridging LiDAR and Camera Modalities for Human Identification". In: *2024 IEEE International Joint Conference on Biometrics (IJCB)*. Sept. 2024, pp. 1–11. DOI: 10.1109/IJCB62174.2024.10744428.

[45] Ting-Chun Wang et al. *Video-to-Video Synthesis*. arXiv:1808.06601 [cs]. Dec. 2018. DOI: 10.48550/arXiv.1808.06601.

[46] Yubo Wang, Bin Liu, Zhiwei Zhao, Jixiang Niu, Qi Chu, and Nenghai Yu. "CMGait: Enhancing Cross-Modality Gait Recognition between LiDAR and RGB through Contrastive Identity-consistent Feature Aggregation". In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889323.

[47] Zhongling Wang, Zhenzhong Chen, and Feng Wu. "Thermal to Visible Facial Image Translation Using Generative Adversarial Networks". In: *IEEE Signal Processing Letters* 25.8 (Aug. 2018), pp. 1161–1165. ISSN: 1558-2361. DOI: 10.1109/LSP.2018.2845692.

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1941-0042. DOI: 10.1109/TIP.2003.819861.

[49] Hongfei Xue et al. "mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave". In: *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. MobiSys '21. New York, NY, USA: Association for Computing Machinery, June 2021, pp. 269–282. ISBN: 978-1-4503-8443-8. DOI: 10.1145/3458864.3467679.

[50] Wei Xue, Hong Ai, Tianyu Sun, Chunfeng Song, Yan Huang, and Liang Wang. "FrameGAN: Increasing the frame rate of gait videos with generative adversarial networks". In: *Neurocomputing* 380 (Mar. 2020), pp. 95–104. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2019.11.015.

[51] Huanqi Yang et al. "XGait: Cross-Modal Translation via Deep Generative Sensing for RF-based Gait Recognition". In: *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. SenSys '23. New York, NY, USA: Association for Computing Machinery, Apr. 2024, pp. 43–55. ISBN: 979-8-4007-0414-7. DOI: 10.1145/3625687.3625792.

[52] Shiqi Yu, Haifeng Chen, Edel B. García Reyes, and Norman Poh. "GaitGAN: Invariant Gait Feature Extraction Using Generative Adversarial Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. July 2017, pp. 532–539. DOI: 10.1109/CVPRW.2017.80.

[53] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. "TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition". In: *2018 International Conference on Biometrics (ICB)*. Feb. 2018, pp. 174–181. DOI: 10.1109/ICB2018.2018.00035.

[54] Peijun Zhao et al. "mID: Tracking and Identifying People with Millimeter Wave Radar". In: *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. May 2019, pp. 33–40. DOI: 10.1109/DCOSS.2019.00028.

# A. Appendix

## A.1. Experiment Results

**Table A.1.:** Subject-level GAN Rank-1 (R@1) and Rank-5 (R@5) accuracies across all short-run splits. Entries are reported only for splits in which the subject appears in the test set; missing entries indicate that the subject was not part of the corresponding test split.

| Subject | Split 0 | | Split 1 | | Split 2 | | Split 3 | | Split 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| 00c24b | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 02a3d4 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 08a482 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 0a96d0 | – | – | 0.2700 | 1.0000 | – | – | – | – | – | – |
| 0b3ffd | – | – | – | – | 0.2700 | 0.6200 | – | – | – | – |
| 0fecb9 | – | – | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.1000 |
| 106a14 | – | – | – | – | 0.3400 | 0.8700 | – | – | – | – |
| 19623a | – | – | 0.0200 | 0.4700 | – | – | – | – | 0.2300 | 0.9500 |
| 19b656 | 0.0000 | 0.7500 | – | – | – | – | – | – | – | – |
| 1a43c5 | – | – | 0.0000 | 0.4900 | 0.0000 | 0.0800 | 0.0000 | 0.0000 | 0.0000 | 0.1700 |
| 1a4c4b | – | – | 0.0000 | 0.0300 | – | – | – | – | – | – |
| 1dd3ca | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – | – | – |
| 1ec949 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 254eb3 | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |
| 259e67 | 0.0300 | 0.5400 | – | – | – | – | – | – | 0.0000 | 0.1100 |
| 26b82c | – | – | – | – | 0.0000 | 0.0000 | – | – | 0.0000 | 0.0000 |
| 2a3d36 | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |
| 2a65f5 | – | – | 0.0000 | 0.7100 | 0.2800 | 0.7500 | – | – | – | – |
| 2bbf07 | – | – | – | – | 0.0000 | 0.1200 | 0.0000 | 0.0000 | – | – |
| 2f4391 | 0.0000 | 0.0000 | 0.0000 | 0.0200 | – | – | – | – | 0.0000 | 0.0000 |
| 30cab2 | 0.5000 | 0.9900 | – | – | 0.7100 | 1.0000 | – | – | – | – |
| 30dff4 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 31b2a3 | – | – | – | – | – | – | – | – | 0.0000 | 0.0000 |
| 32103c | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 344643 | 0.0000 | 0.0000 | 0.0000 | 0.0800 | – | – | – | – | – | – |
| 353e21 | – | – | 0.0300 | 0.1500 | – | – | – | – | 0.0000 | 0.0500 |
| 37772b | – | – | 0.0000 | 0.0700 | – | – | – | – | – | – |

| Subject | Split 0 | | Split 1 | | Split 2 | | Split 3 | | Split 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| 378322 | 0.0000 | 0.0900 | – | – | – | – | – | – | – | – |
| 392867 | – | – | – | – | – | – | – | – | 0.0000 | 0.0100 |
| 39f713 | – | – | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3e5a9f | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3eec0f | – | – | – | – | – | – | 0.0000 | 0.2200 | – | – |
| 4245c8 | – | – | 0.0000 | 0.0000 | – | – | 0.0000 | 0.0000 | – | – |
| 44ef61 | 0.6100 | 1.0000 | – | – | – | – | – | – | 0.1900 | 0.9900 |
| 474df2 | – | – | – | – | 0.0000 | 0.3400 | – | – | – | – |
| 482ff5 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 4a66a4 | – | – | – | – | – | – | – | – | 0.0000 | 0.0100 |
| 4abfd4 | – | – | – | – | – | – | – | – | 0.3500 | 0.8200 |
| 4ca29a | – | – | 0.0000 | 0.0000 | – | – | – | – | 0.0000 | 0.0000 |
| 4db4bd | – | – | – | – | – | – | – | – | 0.0000 | 0.0500 |
| 4db836 | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| 51c25f | – | – | – | – | 0.0100 | 0.4900 | – | – | – | – |
| 539a9a | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – | – | – |
| 546966 | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |
| 5565aa | – | – | – | – | 0.0000 | 0.0000 | – | – | 0.0000 | 0.0000 |
| 560679 | – | – | – | – | 0.2200 | 0.6400 | – | – | – | – |
| 568e2b | – | – | – | – | – | – | – | – | 0.0000 | 0.0000 |
| 56b892 | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – | – | – |
| 57866e | – | – | 0.0000 | 0.0800 | – | – | 0.0000 | 0.0300 | – | – |
| 58d8dc | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |
| 5a4fbd | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| 5bbc8d | – | – | – | – | – | – | 0.4200 | 0.9900 | – | – |
| 5d6423 | 0.0000 | 0.0900 | 0.0000 | 0.1100 | – | – | 0.0300 | 0.0700 | 0.0000 | 0.2100 |
| 5db744 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 5e13f1 | – | – | – | – | – | – | 0.0000 | 0.6200 | – | – |
| 5e66da | – | – | 0.2200 | 0.6100 | 0.0000 | 0.0000 | – | – | – | – |
| 5ea943 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 6259f5 | – | – | – | – | 0.0400 | 0.3100 | – | – | – | – |
| 6313b1 | 0.0000 | 0.1100 | – | – | 0.0400 | 0.2500 | 0.0000 | 0.0900 | – | – |
| 64bed8 | – | – | – | – | 0.0000 | 0.0400 | – | – | – | – |
| 655fc7 | 0.2600 | 0.9000 | – | – | – | – | – | – | – | – |
| 68221d | – | – | 0.0000 | 0.0100 | – | – | – | – | 0.0000 | 0.0000 |
| 68330d | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – | – | – | – | – |
| 69eefe | 0.7700 | 1.0000 | – | – | – | – | – | – | – | – |
| 736784 | 0.0000 | 0.4842 | 0.2211 | 0.5684 | – | – | – | – | – | – |
| 755fa5 | 0.0000 | 0.0571 | – | – | – | – | – | – | – | – |
| 76ad19 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 7986cd | 0.0000 | 0.2300 | – | – | – | – | – | – | – | – |

| Subject | Split 0 | | Split 1 | | Split 2 | | Split 3 | | Split 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| 7a1376 | 0.0000 | 0.0000 | – | – | – | – | – | – | 0.0000 | 0.0100 |
| 7b99ed | – | – | – | – | 0.0200 | 0.4300 | 0.0900 | 0.4900 | – | – |
| 834203 | – | – | – | – | – | – | 0.0000 | 0.2200 | 0.0000 | 0.0000 |
| 85ee9b | – | – | – | – | – | – | 0.0000 | 0.0200 | – | – |
| 865b00 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| 893def | – | – | – | – | 0.1200 | 0.8200 | 0.0100 | 0.9500 | – | – |
| 8d18de | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| 8f14a4 | 0.0000 | 0.1400 | 0.0000 | 0.0600 | – | – | 0.0000 | 0.0400 | – | – |
| 924c85 | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – |
| 9468c7 | – | – | 0.1700 | 0.6400 | – | – | – | – | 0.1000 | 0.5700 |
| 99f4cc | – | – | 0.0000 | 0.0000 | – | – | – | – | 0.0000 | 0.0000 |
| 9cd26d | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |
| 9d2621 | – | – | – | – | 0.0000 | 0.0500 | – | – | – | – |
| 9d2e4e | – | – | 0.0000 | 0.0400 | 0.0000 | 0.4500 | – | – | 0.0200 | 0.2700 |
| 9e141a | 0.0000 | 0.0000 | – | – | 0.2100 | 0.5800 | 0.0200 | 0.1800 | – | – |
| 9e5598 | – | – | – | – | 0.1500 | 0.9100 | – | – | – | – |
| 9fbb54 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| a0d9e0 | 0.3400 | 0.9900 | – | – | – | – | 0.3000 | 0.9300 | 0.2700 | 0.7600 |
| a21c10 | 0.0000 | 0.3300 | 0.0000 | 0.6800 | – | – | 0.0000 | 0.1900 | – | – |
| a3e618 | – | – | – | – | – | – | – | – | 0.3200 | 0.9400 |
| a3f1f6 | – | – | – | – | – | – | 0.5200 | 0.9700 | – | – |
| a534d7 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| a71cff | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – |
| a8b190 | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| a942b5 | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| ab4bea | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| acc825 | – | – | 0.0000 | 0.0000 | – | – | – | – | 0.0300 | 0.1300 |
| aee042 | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| b1d0e1 | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| b4b500 | 0.0000 | 0.3600 | – | – | – | – | – | – | 0.0000 | 0.1100 |
| b86c25 | 0.0000 | 0.0700 | – | – | – | – | – | – | 0.0000 | 0.0100 |
| b8ed4b | – | – | 0.0000 | 0.0000 | – | – | – | – | 0.0000 | 0.0000 |
| bea0a0 | – | – | – | – | 0.0000 | 0.1800 | – | – | – | – |
| c1a735 | 0.0100 | 0.2300 | – | – | – | – | – | – | – | – |
| c5e74a | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0500 | – | – | – | – |
| ca4208 | – | – | – | – | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| ca83dc | 0.0000 | 0.0200 | – | – | – | – | 0.0000 | 0.0200 | – | – |
| cf30ac | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| d42945 | – | – | – | – | – | – | – | – | 0.0000 | 0.1200 |
| dab457 | – | – | – | – | 0.0000 | 0.1600 | – | – | – | – |
| db4a38 | – | – | 0.0000 | 0.1400 | – | – | – | – | – | – |

| Subject | Split 0 | | Split 1 | | Split 2 | | Split 3 | | Split 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| dc32c3 | – | – | 0.0000 | 0.0000 | – | – | – | – | – | – |
| df035b | 0.0000 | 0.0000 | – | – | – | – | – | – | 0.0000 | 0.0000 |
| df1f93 | – | – | 0.0000 | 0.0900 | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| e7b027 | 0.0000 | 0.0000 | – | – | – | – | 0.0000 | 0.0000 | – | – |
| e7e630 | 0.0000 | 0.0000 | – | – | 0.0000 | 0.0000 | 0.0000 | 0.0000 | – | – |
| eae169 | – | – | – | – | – | – | 0.0000 | 0.0000 | – | – |
| eb0ac0 | – | – | – | – | – | – | 0.3200 | 0.9300 | 0.6100 | 0.9800 |
| f14e5f | – | – | – | – | – | – | – | – | 0.0000 | 0.0000 |
| f1b82d | – | – | 0.7400 | 1.0000 | – | – | – | – | – | – |
| f565b4 | 0.0000 | 0.0900 | 0.0100 | 0.1300 | – | – | – | – | – | – |
| f6ad03 | – | – | – | – | 0.0000 | 0.2900 | – | – | – | – |
| fae9ff | – | – | – | – | – | – | – | – | 0.2900 | 0.9100 |
| fbdf5c | 0.0000 | 0.0000 | – | – | – | – | – | – | – | – |
| fcdb7a | – | – | – | – | 0.0000 | 0.0000 | – | – | – | – |

## A.2. Training Details

**Table A.2.:** Concrete values used during training of the proposed radar-to-depth GAN model.

| Attribute | Value |
|---|---|
| Learning rate generator | $1.8981 \cdot 10^{-4}$ |
| Learning rate discriminator | $1.7657 \cdot 10^{-5}$ |
| Adam betas generator | $(0.0, 0.999)$ |
| Adam betas discriminator | $(0.5, 0.999)$ |
| $\lambda_{\text{rec}}$ | 4.8825 |
| $\lambda_{\text{adv}}$ | $2.5951 \cdot 10^{-3}$ |
| $\lambda_{\text{ovl}}$ | 1.2970 |
| $\lambda_{\text{gait}}$ | 3.016 |
| Tversky $\alpha$ | 0.3232 |
| Tversky $\beta$ | 0.6275 |
| Real label for discriminator | 0.95 |