Delft University of Technology

École de technologie supérieure

Master Thesis

Cross-Modal Re-identification of Persons between RGB and Depth

Author:
Frank Hafner

Thesis Committee:
Prof. Dr. Dariu M. Gavrila
Dr. Julian F.P. Kooij
Dr. David Tax
Dr. Wei Pan

Instructors:
Dr. Amran Bhuiyan
Prof. Dr. Éric Granger
Dr. Julian F.P. Kooij

November 17, 2018

Abstract

Cross-modal person re-identification is the task to re-identify a person which was sensed in a first modality, like in visible light (RGB), in a second modality, like depth. Therefore, the challenge is to sense between inputs from separate modalities, without information from both modalities at the same time step. Lately, the scientific challenge of cross-modal person re-identification between depth and RGB is getting more and more attention due to the needs of intelligent vehicles, but also interested parties in the surveillance domain, where sensing in poor illumination is desirable.

Techniques for cross-modal person re-identification have to solve several concurrent tasks. First, techniques have to be robust against variations in the single modalities. Occurring challenges are viewpoint changes, pose variations or variations in camera resolution. Second, the challenge of re-identifying a person has to be solved across the modalities within a heterogeneous network of RGB and depth cameras.

At the present day, work in cross-modal re-identification between infrared images and RGB images exist. At the same time almost no work was done in re-identification between depth images and visible light images. The objective of this work is to fill this gap by comparing the performance of different techniques for cross-modal re-identification of persons. The main contributions of this work are two-fold.

First, different deep neural network architectures for cross-modal re-identification of persons between depth and visible light are investigated and compared.

Second, a new technique for cross-modal person re-identification is presented. The technique is based on two-step cross-distillation and allows to extract similar features from the depth and visible light modality. Therefore, the task of matching persons sensed between depth and visible light is facilitated and can be solved with higher accuracy.

Within the evaluation, it was possible to report state-of-the-art results for two relevant datasets for cross-modal person re-identification between depth and RGB. For the BIWI RGBD-ID dataset the pre-existing state-of-the-art was improved by more than 15% in mean average precision. Additionally, it was possible to validate the performance of the method with the RobotPKU dataset.

Although the method was successfully applied in cross-modal person re-identification between depth and RGB, it was shown that in another modality combinations, like RGB and infrared, the technique in its current definition cannot be considered state-of-the-art.

Finally, it is possible to give a lookout on the implications of the results for the intelligent vehicles domain. For a successful deployment in this area more thorough datasets have to be developed and the performance on sparse depth maps, as provided by lidars or radars, have to be investigated.

Acknowledgements

I would like to express my very great appreciation to Eric Granger, Amran Bhuyian and Julian Kooij for the great support within the emergence of this research. Especially, I want to thank for the great collaboration within the writing of the paper "A Cross-Modal Distillation Network for Person Re-identification between Depth and RGB" which was built upon this thesis. Additionally, I want to thank them for making my research visit in Montreal possible. In this context I also want to thank all students in LIVIA which made my stay exceptional professionally as well as personally.

Furthermore, I want to thank Nicola Schwarz, Daniel Hurst and Felix Heppeler for reviewing my work and Julian Herrmann for kindly providing me with the Latex fonts for PowerPoint.

List of Figures

List of Tables

Contents

1 Introduction

In psychology and neuroscience, cross-modal object re-identification refers to 'the ability to recognize an object, previously inspected with one modality like vision, via a second modality like touch' [1].

For instance, Peter's elephantnose fish is a weakly electric fish which can be found in African freshwater (see figure 1). The fish does not have a complex mammalian brain structure, but it is able to use its vision and its active electric sense for object recognition. In 2016 Schumacher et al. [2] were able to prove that if the fish was trained to discriminate two objects with only one of the two senses, it was subsequently able to succeed in the same task with the other sense. The authors proposed that the fish may have learned low-level features to associate electric and visual input through analyzing other environmental objects in the past. In fact, the fish is capable of cross-modal object re-identification, despite its simple brain structures [2].

Figure 1: Peter's elephantnose fish [3].

A lot of current research in computer vision is performed in mutual evaluation of two or more modalities, like object detection with RGB and depth (RGB-D) information. Nevertheless, in the future the importance of performing visual tasks across modalities will rise.

In this work a specific cross-modal task will be investigated, which is cross-modal person re-identification between sensors with focus on RGB and depth inputs. The problem is defined as having a gallery person image from one modality, like a visible light image, and having query person images from another modality, like depth images. The task is to correctly match query and gallery images from the same person in a defined search space. An example for this setup can be seen in figure 2.

The task of re-identification of persons in a single modality was investigated a lot in recent years. The main challenges are pose differences, lighting variations and camera resolution differences within images of a single person instance. For the cross-modal task additionally a common sensing between two heterogeneous sensor modalities is needed. This adds another challenge to the task and underlines the need for robust techniques for the specific challenge.

In surveillance, re-identification of an object or person is crucial as a consistent observation through its occurrence in a relevant area is desired. Therefore, several often non-overlapping sensors have to be analyzed such that the same object can be re-identified reliably. Nowadays, these Closed Circuit Television (CCTV) systems are mostly realized with several cameras sensing in the RGB modality. The re-identification task within a CCTV with visible light cameras involves challenging person re-identification situations. An excerpt of those situations are visualized in figure 3. From left to right the challenges

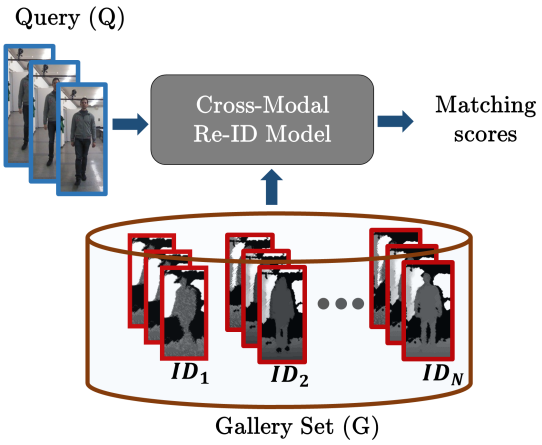


Figure 2: Illustration of the cross-modal person re-identification system based on RGB (query) and depth (gallery set) modalities.

of low resolution, occlusion, viewpoint changes, pose and illumination variations and similar appearance of different person instances are shown.

Zhuo et al. [5] describe that these systems additionally have problems in dark environments, since appearance features are not sensible with visible light cameras in these surroundings. Therefore, a potential solution can be to use a depth capturing device in dark environments, while still relying on RGB cameras in light settings [5]. Especially the recent progress in lidar technology makes depth measurements a feasible alternative to infrared cameras in these cases [6, 7].

Also in environment sensing of intelligent vehicles, cross-modal re-identification of persons is relevant. A typical design of a sensor set for an autonomous car can be seen in figure 4. In this example, none of the sensors is taking into account a full 360 degree view. For example a camera sensor is only used for frontal view, whereas lidars cover the front as well as the sides of the car. Hence, sensor information which is available for different views from the car is a combination of one or more modalities with another combination of modalities. To get a thorough understanding of the environment, the vehicles seeks to be able to sense between sensor modalities to minimize uncertainties.

An exemplary situation where cross-modal re-identification can can be beneficial for the environment sensing of a vehicle can be seen in figure 5. Here, the vehicle is equipped with a front-facing camera and a side-facing lidar. An obstacle avoids that the sensors are

Figure 3: Challenges in Person Re-identification (from left to right): low resolution, occlusion, viewpoint changes, pose and illumination variations and similar appearance of different people [4].

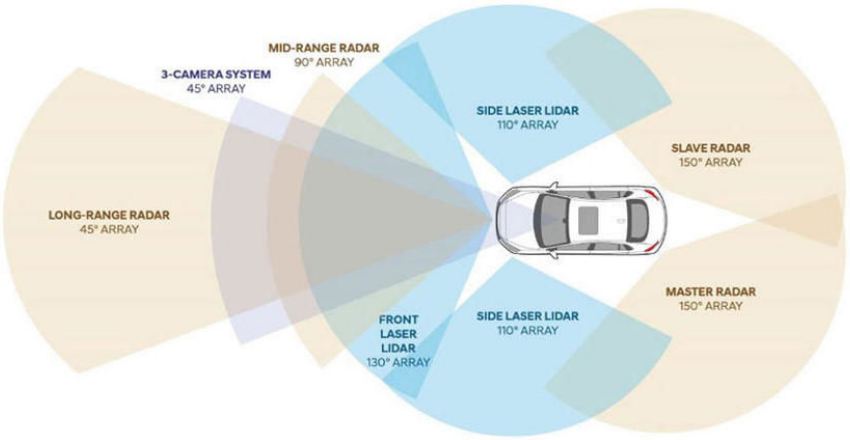


Figure 4: Examples for a sensor set for autonomous driving from Hyundai [8].

overlapping. In the scene two objects, e.g. pedestrians, are moving in the environment of the car. In time step t the green object is visible in the camera domain and no object is visible in the lidar domain. In time step $t+1$ one object is visible in the lidar domain and none in the camera domain. To identify the red object in the lidar domain as a new object gives several advantages. First, it is known, that the green object is still behind the obstacle and has to be considered when taking a left turn. Second, assuming a movement model for the objects, it is known that there is no prior information about the movement patterns of the red object, as it was not seen before.

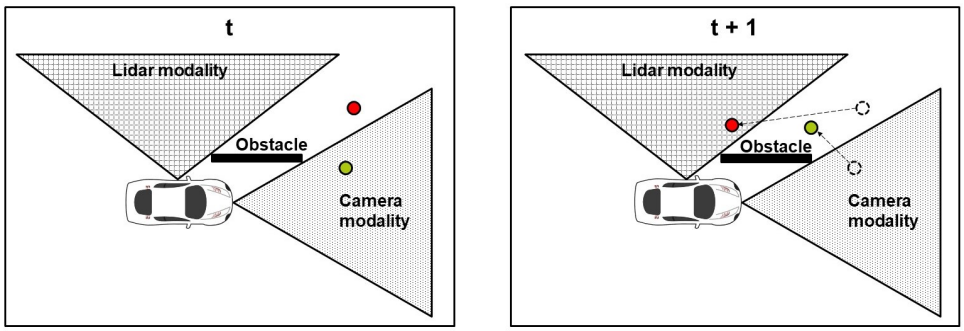


Figure 5: Left: Situation at time step t; Right: Situation at time step t+ 1. The green and red dots are moving objects.

The **research questions** of the thesis can be defined as follows:

"How and how well can the cross-modal person re-identification task between RGB and depth be solved?"

This question is complemented by an additional research questions:

"Is it possible to develop generic techniques for person re-identification between modal-

ities and how well do they generalize to different modality combinations, like infrared images and visible light images?"

To answer these questions in chapter 2 related work on the topic is presented. Following the most relevant approaches in the found literature, in chapter 3 three neural network techniques for solving the task of cross-modal person re-identification will be presented. Two of those techniques are extracted from related literature on RGB-infrared re-identification and one technique is a contribution of this work. In chapter 4 relevant datasets and evaluation measures which are vital for the understanding of this work will be discussed and, subsequently, experimental details will be analyzed. In chapter 5 results of the experiments in this work will be investigated. Therefore, in chapter 5.1 networks for single-modal re-identification will be presented and evaluated. This chapter acts as the baseline and comparison point for the subsequent chapters. Chapter 3 is the main part of this work. Here, the presented neural network methods will be presented and evaluated. In chapter 5.3 the methods will be compared and placed in a wider context by comparing the results with external work. Finally, in chapter 6 conclusions on the results will be made. Based on this, an outlook on the implications for the intelligent vehicles community and future work will be given.

2 Related work

As not a lot of work concerning the task of cross-modal re-identification of persons exists, it is necessary to take a look at related subject areas to obtain a full understanding for the challenges and its solutions space. Therefore, first, an introduction to the general single-modal re-identification task will be given, which will be extended with an introduction to cross-modal re-identification (section 2.1). After this, a look at the literature for re-identification in single modalities will be taken in chapter 2.2. Therefore, an condensed overview of techniques for person re-identification in single-modalities will be given in section 2.2.1. The analysis shows that deep neural networks are the current state-of-the-art for single-modal sensing. Therefore, a look at common losses and architectures for training of neural networks will be taken in section 2.2.2.

Finally, the literature for cross-modal re-identification will be investigated in section 2.3. Therefore, the task of cross-modal re-identification will be interpreted as an domain adaptation task and the solution space will be discussed (section 2.3.1). Furthermore, existing literature for cross-modal re-identification between depth and visible light images of persons will be discussed in section 2.3.2 and existing literature from cross-modal person re-identification between infrared and visible light, which got more attention from the scientific community recently will be investigated 2.3.3.

2.1 The re-identification task

Gong et al. [9] describe a re-identification task metaphorically as, firstly, 'finding needles in haystacks' and, secondly, 'connecting the dots'. What they pin as 'finding the needles' is the scientific challenge of object detection. For this task many successful approaches exist. Before 2012 these methods were mainly based on hand-crafted features, like Haar Cascades for face detection [10] or contour and intensity features for pedestrians [11]. Since around 2012 deep-learning based methods like Faster R-CNN or R-FCN are more successful and mostly the state-of-the-art in the area [12, 13]. For the re-identification task in this work no deeper insights into object detection algorithms will be given and the objects of interest are assumed to be detected. For further insights into object detection please refer to the mentioned literature.

For the re-identification task, most methods are based on bounding boxes. However, to minimize the influence of background clutters it can be beneficial to have a more exact instance segmentation. For this work, bounding box labels are assumed to be provided for all datasets. Starting from this point Gong et al. define the re-identification pipeline for computer vision as follows [9]:

1. Extraction of features which are descriptive from the raw pixel input.
2. Construction of a descriptor or a representation based on the extracted features.
3. Definition of probe and gallery images and matching of those by hands of the descriptors or representations.

A visualization of this pipeline in a single-modal task can be seen in figure 6. In a single-modal task the re-identification pipeline starts in the same input space χ which is, in the given example, an RGB colour input of persons. Given are a query image $\hat{x}$ with label $\hat{y}$ and a set of gallery images $x_1, \cdots, x_M$ with labels $y_1, \cdots, y_M$. Query as well as

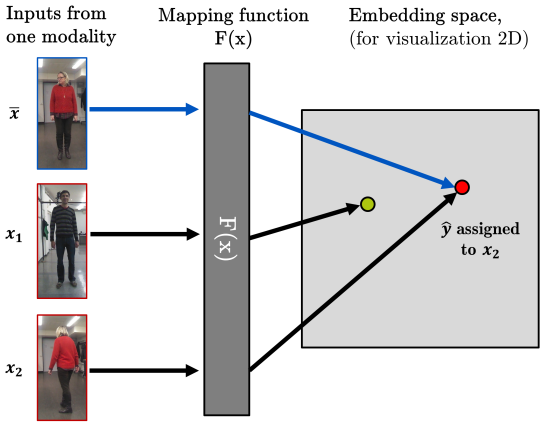


Figure 6: Single-modal re-identification: Embedding from the same input to a common feature space. Top and bottom image on the left indicating the same person.

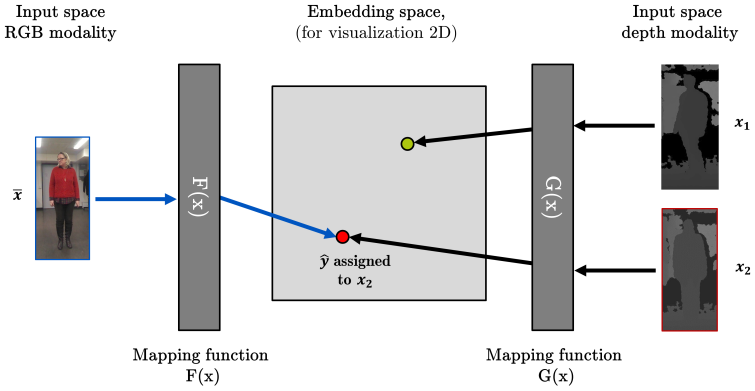


Figure 7: Cross-modal re-identification: Embedding from different input spaces to a common feature space. Left and right bottom from the same person.

gallery images are mapped to a latent embedded space with a function $F(x)$. Ideally, the function $F(x)$ is designed such that the embedding $F(x_i)$ of image x_i with the same label as the query $\hat{y} = y_i$ is close to the embedding $F(\hat{x})$ of query image $\hat{x}$ in the latent space and the embedding $F(x_i)$ of image x_i with a different label as the query $\hat{y} \neq y_i$ is far in the latent space.

The label of the corresponding person in the gallery for query $\hat{x}$ is then assigned to the individual corresponding to the closest embedded gallery image $\hat{i}$, i.e.

$$\hat{y} = y_{\hat{i}} \qquad \text{where} \qquad \hat{i} = \arg\min_i d(F(\hat{x}), F(x_i)). \tag{1}$$

where d is the distance metric for the embedding. In general different distance metrics are applicable. In the exemplary visualization in figure 7, the deployed metric is Euclidean distance. In single-modal re-identification, both query $\hat{x} \in \chi$ and gallery images $x_i \in \chi$ are from the same input space χ.

In figure 7 a scheme for cross-modal re-identification is given. The main difference to the single-modal task is, that query $\hat{x} \in \chi_s$ and gallery images $x_i \in \chi_t$ are extracted from

a different feature space. Therefore, it is necessary to define a second mapping function $G(x)$ which maps to the same latent embedded space as $F(x)$, to be able to find $\hat{i}$. In the scheme (see figure 7) the query image is from the RGB modality and the gallery images are from the depth modality. However, the assignment of query and gallery to the modalities depth and RGB is to be varied.

In general, the core difference of a re-identification task to a standard classification task in machine learning is that classes, like e.g. specific persons, which have to be classified at test time, are not part of the training set. Therefore, re-identification tasks can be defined as zero-shot learning tasks [14].

2.2 Techniques for re-identification in single modalities

Research for re-identification of different objects was advancing heavily in recent years and a lot of scientific work was published [15, 16, 17, 18]. It turned out that specific neural network architectures and training schemes are in the core applicable for several re-identification tasks in computer vision, like faces, person or vehicles in a similar manner. As the focus of this work is person re-identification an overview of scientific work in re-identification of persons in single modalities will be presented in the following.

It turns out, that the current state-of-the-art in person re-identification are deep neural network. As these techniques will be used in the latter parts of this work in the following several neural network ingredients which are proven to be applicable for person re-identification will be presented. Hence, it is no exhaustive discussion of neural network building blocks, but very specific to the case of re-identification. For a more detailed overview on definitions and mathematical formulations for neural networks in general please refer to [19].

2.2.1 Person re-identification

As the focus of this work is on person re-identification the most important ideas for this task in single modalities will be presented in the following. The biggest challenges in Person re-identification are the problem of different viewpoints, changing orientation of persons and changing lighting conditions. Additionally, especially in long-termed tasks, appearance changes through different clothing is a challenging characteristic of person re-identification [20].

In this chapter, first, a view on conventional approaches for person re-identification will be discussed. Afterwards, ideas for neural network based approaches will be shown. Additionally, a separate view on the depth domain will be taken. The most recent survey for single-modal person re-identification dates to 2016 and can be found in [20].

2.2.1.1 Conventional approaches

Conventional approaches for person re-identification from a single modality can be categorized into two main groups - direct methods based on hand-crafted descriptors or learned features and metric learning based approaches. Direct methods for re-identification are mainly devoted to the search of the most discriminant features, or combinations thereof, to design a powerful descriptor or signature for each individual regardless of the scene. In contrast, in metric learning methods, a dataset of different labeled individuals

is used to jointly learn the features and the metric space to compare them, in order to guarantee a high re-identification rate.

Due to the non-rigid structure of the human body, it is difficult to model the appearance of the whole body for re-identification. Instead it is more robust to model the appearance focusing on salient parts or meaningful parts of the body. Most of the direct method based re-identification approaches rely on the local meaningful parts, e.g. horizontal stripes [21, 22], triangular graphs, concentric rings [23], symmetry-driven structures [24], pictorial structure [25], meaningful body-parts [26] and horizontal patches [27]. Different features such as color based features [28, 26], textures [29, 30, 31], edges [31], Haar-like features [32], interest points [33] and Biologically Inspired Features (BIF) [33] and different combination of those features such as Bag-of-Words (BoW) [34], Weighted Histogram of Overlapping Strips (WHOS) [35], and Local Maximal Occurrence (LOMO) [21] from those local regions have proven to be useful to achieve better re-identification accuracy. Given the handcrafted features, another stream of direct method based re-identification approaches learns the feature importance based on the salient feature analysis of each individual [36, 37, 26] , or by exploiting the coherence among different features on a manifold space [38].

Metric learning based approaches usually find a mapping from feature space to a new space in which feature vectors from image pairs of the same individual are closer than feature vectors from different image pairs. Commonly used metric learning techniques that are adopted for re-identification include Mahalanobis metric learning [39], Large Margin Nearest Neighbor Learning (LMNN) [60], Logistic Discriminant Metric Learning (LDML) [60], Kernel Canonical Correlation Analysis (KCCA) [46], keep it simple and straight forward metric learning (KISSME) [39] and Cross-view Quadratic Discriminant Analysis (XQDA) [21] .

2.2.1.2 Deep Learning Methods

Similar to other vision applications, there has been a growing number of deep learning based re-identification approaches [40, 41, 42, 43, 44, 45, 49, 48]. One stream of works for person re-identification is using the ideas from Siamese CNN with either two [40, 41, 42, 48, 43, 44] or three branches [45, 49, 50] for pairwise verification loss or combination of both [51]. Another stream of works is based on softmax loss to obtain an generalized embedding layer [52, 53]. These losses will are considered fundamental ingredients to neural networks for re-identification and will be presented in more detail in chapter 2.2.2.2. Some of those approaches use their own network architectures, by proposing new layers [41] or by fusing features from different body parts with a multi-scale CNN structure [42, 54]. Some other [45, 55, 50] use the pre-trained or different variants of pre-trained models (e.g. Resnet [56]) which often obtain great re-identification accuracy.

2.2.1.3 Incorporating depth

Several works [61, 59], have identified the fact that theoretically methods based purely on RGB appearance can be problematic for re-identification purposes due to changing appearance of pedestrians. Therefore, a solution can be to use depth information for the classification.

The solutions which are incorporating depth can be divided into two approaches. Those solely relying on one image [62, 63] and those taking into account several images to leverage spatio-temporal behaviour [64, 65, 66, 67].

Early single-image depth-based studies relied on the extraction of anthropometric and soft-biometrics from 3D human skeleton [62, 63]. Other approaches were analyzing the 3D point-clouds of humans by hand-crafted features, like arm length and torso width or by RGB based features like SIFT or SURF [64]. Other approaches were built upon incorporating spatio-temporal information for the re-identification. Already in the mid-2000ers several researchers tried to extract gait information from pedestrians, leveraging the skeleton information which can be extracted by hands of a Kinect camera. From the given skeleton information over time different hand-crafted features were extracted and matched by techniques, like k-Nearest Neighbor. It was shown, that gait is unique for each person [64, 65].

Due to the success of the methods, researchers in the deep learning era incorporated the gait information within recurrent neural networks. Karianakis et al. [66] presented a recurrent neural network for the re-identification task, which is based on the features of a 3D input convolutional neural network. These features are combined by a recurrent neural network which weights the input frames with a temporal attention unit technique [66]. In contrast, Haque et al. [67] formulate an attention-based recurrent neural network which identifies small discriminative regions of a human body for describing a human identity [67].

2.2.2 Training of feature extractors for re-identification

In chapter 2.2.1 it was shown, that current approaches for re-identification are mainly built upon deep neural networks. Historically, algorithms for re-identification were mainly developed for face re-identification. Hardly surprising, with the rise of neural networks the most important breakthroughs were achieved in this area. Nowadays, face re-identification surpassed the human performance and the successful techniques are transferred to other re-identification tasks, like person or vehicle re-identification [68, 71, 72]. In fact, several building blocks for neural networks for re-identification have proven to be applicable for several of the tasks. Also most of literature found in section 2.2.1.2 relies on these building blocks. Therefore, in the following a feature extractor architecture with two variants and several loss functions for training will be presented. These are the building blocks for the neural networks in the later chapters.

The overview of convolutional neural networks and corresponding losses in the following is not exhaustive and crafted to the applications later in this work. For further information please refer to [19]. In the following a feature extractor architecture based on a convolutional neural network as well as two loss functions are presented.

2.2.2.1 CNN Feature Extractor: Residual network (Resnet)

One of the most successful convolutional neural network architectures for feature extraction are residual networks. Soon after the first success of convolutional neural networks the scientific community realized, that it was necessary to have deeper networks to avoid overfitting to datasets. However, stacking more layers in a network led to the vanishing gradient problem. A vanishing gradient occurs, when back-propagating through a lot of layers with repeated multiplications. This makes the gradient infinitesimally small and meaningful learning is not possible anymore.

The core idea of residual networks which was introduced by He et al. [56] is to use identity shortcut connections to skip layers in the network. An example for such a skip connection can be seen in image 8. The connections are applied in several layers and

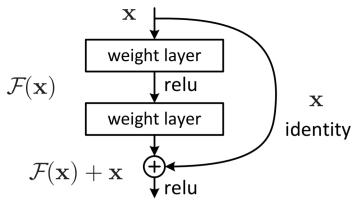


Figure 8: Example for an residual layer [56].

lead to the capability of training much deeper neural networks successfully. In this work mainly a more shallow Resnet with 18 layers (Resnet18) and a deeper Resnet with 50 layers (Resnet50) will be considered to analyze the effect of varying the depth of networks. The exact structure of the two networks can be seen in figure 9. Residual layers are introduced around each block in the figure. Additionally, He et al [56] presented Resnets with 34, 101 and 152 layers. For the sake of compactness of this work, those will not be further considered.

layer name	output size	18-layer	50-layer
conv1	112×112	7×7, 64, stride 2	
conv2_x	56×56	3×3 max pool, stride 2	
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax	
FLOPs		1.8×10^9	3.8×10^9

Figure 9: Structure of Resnet18 and Resnet50 residual layers are introduced around each block of two layers [56].

2.2.2.2 Softmax Loss

The first loss to be presented is the *softmax loss* as used in [73]. During training the softmax loss optimizes the probability of training images $x_1, \cdots, x_M$ to belong to a corresponding label $y_1, \cdots, y_M$, where C different classes are existent in the training set.

Softmax loss is generally used for classification tasks and, therefore, has to be re-interpreted for usage in zero-shot learning for re-identification. Instead of using the probability outputs of the softmax layer as in classification tasks, for the usage in re-identification a preceding layer has to be interpreted as an embedding layer. The idea is that this embedding layer optimized with many person instances generalizes to generic features. Therefore, this embedding layer enables to compare person instances which were not part of the training set and, hence, can be used for zero-shot learning.

In most previous work the penultimate layer before the softmax classification layer is used as the embedding layer [73, 74]. Here, the embedding is defined as $F(x_i) = z_1, \cdots, z_n$ where n is the size of the embedding layer which can be chosen freely. Therefore, the

softmax loss is defined as

$$L_{soft1} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{W_{(y_i)}F(x_i)+b}}{\sum_{j=1}^{C}e^{W_{(j)}F(x_i)+b}}\right), \tag{2}$$

where W are the weights and b is the bias of the penultimate layer. The subscripts of W indicates the subset of weights which map to the corresponding embedding feature of the classification nodes. N corresponds to the batch size.

In this work an additional way of extracting an embedding from a network optimized with the softmax loss will be investigated. Here, the values of the classification layer before fed into the softmax loss are defined as the embedding. Therefore, the embedding is defined as $F(x_i) = z_1, \cdots, z_C$. Due to its definition, this embedding can only have the size of the classes in the training set C. The softmax loss for this interpretation is defined as

$$L_{soft2} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{z_i^{(y_i)}}}{\sum_{j=1}^{C}e^{z_i^{(j)}}}\right). \tag{3}$$

The embeddings obtained from the softmax are not inherently trained to be optimized within a certain metric. Hence, different metrics like L_1 or L_2, or even a second metric learning step can be deployed. To reduce the complexity in this work, in the following solely the Euclidean distance will be evaluated as a metric. Also further ideas of enhancing the performance of softmax loss, like center loss [72] will not be further investigated.

2.2.2.3 Metric Losses

Already in 2005, Chopra et al. [75] presented a *Siamese neural network* structure which is optimized by hands of the *contrastive loss*. The general idea of mapping input images into common subspaces, e.g. in a space for Euclidean loss, is described in this paper. The contrastive loss is based on comparing the Euclidean distance between the embeddings $F(x_i)$ and $F(x_j)$ of two images x_i and x_j. The loss seeks to find an embedding of the images such that a small distance is obtained for the embeddings of two images of the same class and a big distance for the embeddings of two images of different classes. For both images the function F with shared parameters W is used and, hence, the network is evaluated in parallel. Therefore, theoretically, the network consists of two identical networks with one cost function, which makes it 'Siamese' [75]. The last layer which is directly optimized through the loss is considered the embedding for the face.

In 2015 Schroff et al. [68] presented an expansion to the contrastive loss, the *triplet loss*. A big point of criticism for the use of softmax for obtaining the features was, that it only encourages the separability of features for seen objects and, therefore, is indirect and inefficient for unseen objects. They even state, that "one has to hope that the bottleneck representation generalizes well to new objects" and, additionally, criticize the high-dimensionality of the embedding features of most papers at that point in time. For the triplet loss at each training step three images are evaluated, an anchor image x_i^a, a positive image x_i^p and a negative image x_i^n. Anchor and positive are images of the same person, while the negative corresponds to a different person. Defining $F(x)$ as the embedding function the goal is to match the constraint

$$d(F(x_i^a) - F(x_i^p)) + \alpha < d(F(x_i^a) - F(x_i^n)). \tag{4}$$

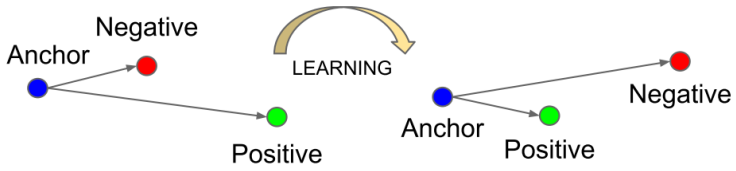


Figure 10: Illustration of the triplet loss [68].

This means the distance d of embeddings of two images of the same person has to be smaller than the distance of the embedding of the anchor to the image of the other person by a margin α. Therefore, the authors argue that the loss enforces a margin between each pair of faces, which should lead to a better embedding. The optimization constraint is also illustrated in figure 10. The constraint can be reformulated as a loss function as

$$L_{trip} = -\frac{1}{M}\sum_{i=1}^{M}[d(F(x_i^a), F(x_i^p)|) - d(F(x_i^a) - F(x_i^n)) + \alpha] \tag{5}$$

where M is the number of parallel evaluations of the triplet loss within one batch. In most publications the Euclidean distance is taken as the measure for d.

Additionally, several authors suggest that some effort has to be taken to choose the right triplets for training [75, 45]. Too easy triplets lead to very slow convergence and very difficult triplets lead to numerical instability. In this work, the most difficult triplets within each batch will be used for training. This method, does not find the most difficult triplets over the whole training set, but semi-hard triplets. Hence, the training is more stable. The main advantage of using the triplet loss is, that the last feature layer is directly optimized to develop features which are separable in Euclidean space. Hence, trained with a sufficiently big dataset a network optimized with the triplet loss is inherently capable of embedding and re-identifying previously unseen objects.

2.3 Techniques for cross-modal re-identification

Whilst the scientific community focused heavily on person re-identification in a single modality (see previous chapter), the task of cross-modal re-identification, which is the actual challenge approached in this work, was mainly neglected. Cross-modal re-identification can be classified as a domain adaptation task. In chapter 2.3.1 an analysis of how the task can be defined in this context and which possibilities this definition introduces will be given. Additionally, very recent work was published on cross-modal re-identification of persons between RGB and depth as well as RGB and infrared. This work will be presented in the following.

2.3.1 Re-identification as domain adaptation

To classify the task of cross-modal re-identification in the context of domain adaptation it is necessary to take a look at several definitions and notations in the domain. D^s is defined as the source domain, while D^t is the target domain. A domain D consists of a feature space χ and a probability distribution $P(\chi)$. To each domain belongs a task, like object detection or re-identification. The task in the source domain is noted as T^s, while the task in the target domain is denoted T^t. In traditional machine learning $D^s = D^t$ and

$T^s = T^t$ holds. Therefore, domains and tasks are the same. The definition for transfer learning is, that either $D^s \neq D^t$ and $T^s \neq T^t$.

Finally, domain adaptation is defined as $T^s = T^t$ and $D^s \neq D^t$. Following the definition for D, there are two categories for domain adaptation. These are homogeneous and heterogeneous domain adaptation. Homogeneous domain adaptation is defined as an identical feature space $\chi^s = \chi^t$ with a difference in the data distributions $P(X)^s = P(X)^t$. On the other hand heterogeneous domain adaptation is defined as non-equivalence of features spaces $\chi^s \neq \chi^t$. Additionally, the dimensions of the feature space can differ. Homogeneous as well as heterogeneous domain adaptation can be subdivided in supervised, semi-supervised and unsupervised domain adaptation.

Cross-modal person re-identification is a domain transfer task in two different domains $D^s \neq D^t$ and with a common re-identification task $T^s = T^t$. Additionally, the feature spaces are different, as it handles two different modalities and, therefore, $\chi^s \neq \chi^t$. As labels are available it is a supervised domain adaption. Hence, cross-modal re-identification can be defined as supervised heterogeneous domain adaptation [76]. However, the cross-modal task is to sense across two domains and, therefore, solve the tasks T^s and T^t in a common space. In general, the goal of domain adaptation is to solve task T^t in the target domain and use the knowledge obtained in the source domain. Therefore, the general definition is not asking for a common solution of both tasks and cross-modal re-identification is even more complex than general heterogeneous domain adaptation.

Wang et al. [76] find, that not much work was focused on heterogeneous domain adaptation so far and even state, that "special and effective methods of heterogeneous domain adaptation have not been proposed." [76]. The existing methods are mostly performed similar to approaches for homogeneous domain adaptation. The implication for this work is, that there is a need for new techniques for heterogeneous domain adaptation which make use of the specific properties of the contained modalities. In chapter 5.2.3 a new technique will be presented.

As existing solutions the authors define adversarial approaches, reconstruction-based approaches and discrepancy-based approaches. Adversarial approaches are mainly focused on unsupervised tasks like transferring knowledge from unlabeled face images to sketches. Reconstruction-based approaches are also based on generative adversarial networks (GAN) for reconstruction of the two different domains.

For discrepancy-based approaches a work was found which is relevant for this work as it is concerned with sensing between RGB and depth. Gupta et al. [77] presented a method for "Cross Modal Distillation for Supervision Transfer". Their goal is to use learned representations from large datasets in a certain modality for classification in a paired modality with limited labeled data. An example usage of the authors is transferring the capabilities of a CNN object classificator in RGB to the corresponding depth images. The method is based on the availability of large amounts of unlabeled coupled images from both modalities.

The modality with a lot of labeled data is defined as D_s whereas the modality with few labeled data is D_t. The corresponding mapping functions (in this case mostly neural networks) are $F(x)$ and $G(x)$. The authors propose a learning scheme where the mapping output of a mid-level layer of each image $x_{i,m1}$ from modality D_s is supposed to match the mapping for the coupled image $x_{i,m2}$ from D_t (see figure 11). Therefore, a mid-level layer is fixed for the optimization and the previous layers are optimized with the unlabeled training data from the second domain. The optimization itself is achieved by a mean squared error loss between the two modalities to minimize the Euclidean distance. This approach is used as a pre-training procedure. After the mid-level layers converged to a

similar embedding for both modalities, the mid- to high-level layers are unfrozen and the network is trained as in a single-modal task.

The authors find that mid-level layers of a network are best suited for freezing and learning the transfer.

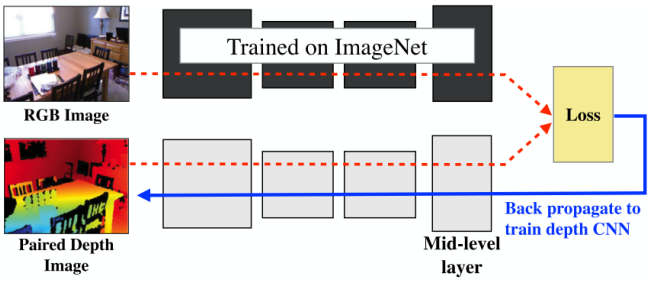


Figure 11: Transfer learning scheme in Gupta et al. [77].

The described training procedure gives a starting point for connecting the modalities depth and RGB. The method which will be developed in this work will expand this idea to make it usable for cross-modal re-identification.

2.3.2 Re-Identification in RGB-Depth

The work which was done on cross-modal re-identification in RGB and depth is very sparse. In fact only one stream of works was found which is connected to cross-modal person re-identification in these modalities.

In 2017 Wu et al. [78] proposed a depth-shape descriptor called *eigen-depth* to extract describing features from the depth domain. The eigen-depth features are based on a division of the body into several describing regions or voxels and an extraction of within voxel and between voxel covariances (see figure 12). The eigenvalues of these covariance matrices are logarithmized and used as the eigen-depth features. The distance between eigen-depth features are proven to lie in Euclidean space and are rotation invariant. The authors were able to show, that those orientation-invariant descriptors of body regions are less prone to errors due to position and lighting changes. As their result in the depth domain were very promising, they decided to transfer the obtained knowledge to the RGB domain. The argumentation is that for most surveillance cameras no depth

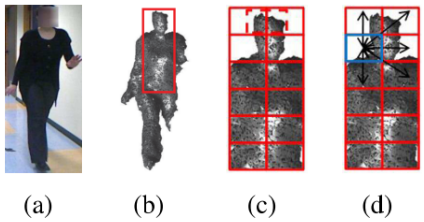


Figure 12: Extraction dimensions for the eigen-depth. (a) RGB image; (b) depth point cloud; (c) within voxel feature extraction; (d) between voxel feature extraction [78].

information can be extracted, but still the features captured in the depth domain are more discriminative than the ones from the RGB domain. Therefore, the authors extract Histogram of Gradients (HoG) [79] and LBP features [80] from the RGB domain, as these

features are supposed to describe the body shape coarsely. The goal is to learn a common subspace representation by hands of mappings $F(x)$ for RGB and $G(x)$ for depth, from the features extracted in the different domains. To achieve this goal, the authors define an optimization problem which can be solved with an Eigendecomposition. Finally, a common latent subspace can be defined. The authors use the obtained transformation to bring the extracted features from RGB to the subspace and perform the re-identification task in this subspace. The scheme of the idea of the authors can be seen in figure 13. Although, the methodology is in principle applicable in cross-modal re-identification, the authors do not perform any evaluations for cross-modal re-identification [59].

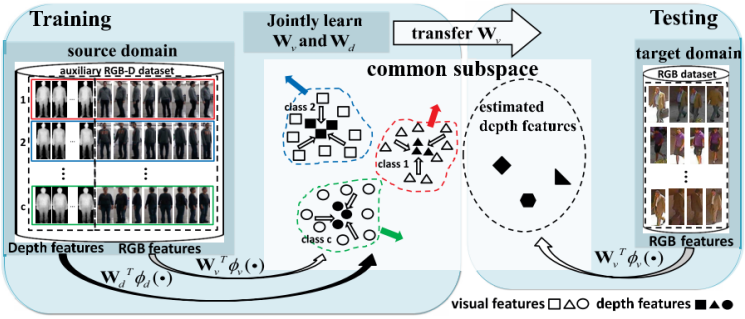


Figure 13: Scheme in Wu et al. [78].

Zhuo et al. [5] take the ideas presented in the paper by Wu et al. [59] and use them for the cross-modal re-identification task. In the paper "Person Re-identification on Heterogeneous Camera Network" [5] they use hand-crafted features for extracting descriptors for RGB and depth images. For RGB they make use of Histogram of Gradient (HoG) [79] and Scale-Invariant Local Ternary Patterns (SILTP) [81] features. For the depth image Eigen-depth features [78] are used. Similarly to Wu et al. [59] they argue that the extracted features for both modalities describe human body shape and, therefore, are inherently already reducing the discrepancy between the modalities. Nevertheless, this similarity is by far not enough to directly compare the extracted features and another step has to be taken to match the spaces. The authors propose the learning of a coupled dictionary for matching the features. This technique is based on an optimization of a convex problem and delivers correlative dictionaries. The obtained sparse vectors for both input modalities can now be compared in Euclidean space.

After the presentation and optimization of the cross-modal re-identification techniques in this paper, a competitive evaluation will be made in chapter 5.3, where the results of Zhuo et al. will be discussed.

2.3.3 Re-identification in RGB-Infrared

For cross-modal person re-identification between RGB and Infrared recent work is available. In contrast to person re-identification between RGB and depth in these domains, neural network techniques are already used. In the following the relevant papers will be discussed.

In 2017 Wu et al. published the paper "RGB-Infrared Cross-Modality Person Re-Identification" [83] were they presented the SYSU-IR dataset. The datasets consists of RGB and infrared images and was developed for cross-modality re-identification of persons. For more details on the dataset refer to chapter 4.1.3. Also cross-modal re-identification between infrared and RGB images is motivated by surveillance applications. The authors

analyze several standard neural network structures to embed the two modalities to one. First, a one-stream neural network which simply takes mixed inputs from the modalities as equally weighted is presented (see also chapter 5.2.1). Second, a two-stream neural network, which gives the network two input streams which are evaluated separately first and connected in a subsequent layer. Third, they evaluate a newly developed network, which they call "One stream structure with zero-padding augmentation" network. The idea is to define a network with two input channels, one for each modality. Therefore, if the input is from the one modality, the channel of the other modality is padded with zeros. The approach can be seen in figure 14. With this method the authors give the network a guideline on specific nodes for the first modality, specific nodes for the second modality and shared nodes, but also the possibility to freely combine the nodes. With this method the authors set the state-of-the-art in the SYSU dataset (see chapter 4.1.3).

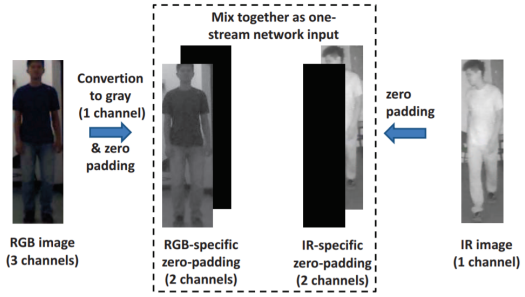


Figure 14: Input manipulation for zero-padding [83].

In mid-2018 Ye et al. [84] published two papers connected to the topic of cross-modal person re-identification in RGB vs. Infrared. In the first one [84], they presented a two-stream neural network which combined a contrastive and a softmax loss together. To enhance the results they attached a subsequent metric learning step. The results were only reported on a dataset which was not made available for this work.

The second work [85] is more relevant, as the results were published on the SYSU dataset. Here, the authors adopt the same methodology as used in [84] and combine two losses. The first loss has the goal to minimize the cross-modal intra-distance and at the same time maximize the inter-modal distances. Hence, the authors compare the distance of a positive visible-thermal image pair and the minimum distance of all negative visible-thermal pairs. This is very much related to a standard triplet loss. This loss is accompanied by an identity loss to guarantee the robustness (see chapter 4.3.2 for more details on the difficulties with robustness in triplet loss). The authors manage to enhance the performance on SYSU. The network structure can be seen in figure 15.

The current state-of-the-art in SYSU was published by Dai et al. [86] in July 2018 under sthe title "Cross-Modality Person Re-Identification with Generative Adversarial Training". The idea of the authors was to combine three losses. The first two losses are a softmax loss and a triplet loss. Therefore, they combine two of the methods, which are in general separately capable of training a neural network in one modality (see explanation in chapter 2.2.2). Additionally, they introduce a GAN based structure. The discriminator differentiates from which modality the input sample came and, hence, the generator enforces a mutual embedding. With this method the authors managed to push the re-identification performance on the SYSU dataset significantly by 11%.

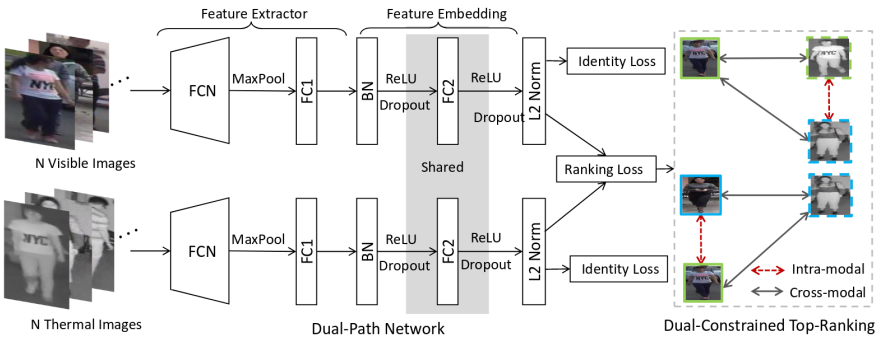


Figure 15: Two-stream network as defined by Ye et al. [85].

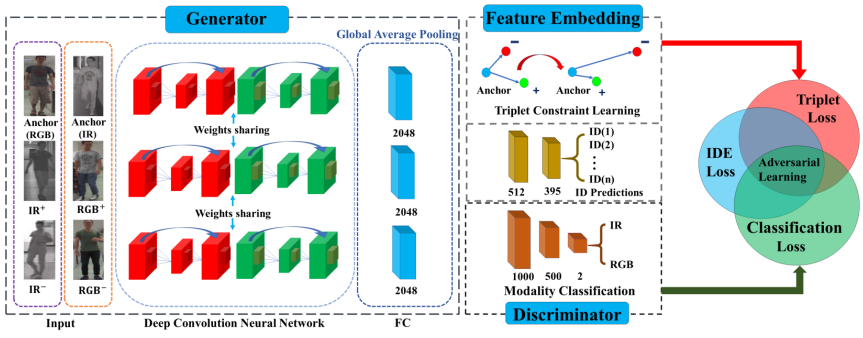


Figure 16: Scheme of Generative Adversarial Training for cross-modal re-identification [86].

2.4 Main contributions

After a thorough analysis of the related work for cross-modal re-identification it is possible to classify the main contributions of this work in this context.

First, deep neural networks are successfully deployed for the task of cross-modal person re-identification between depth and RGB. To the present day only conventional methods were used to solve this task. In this work two deep learning methods from the related task of person re-identification between infrared and RGB will be presented and used for a classification in the re-identification task between RGB and depth. Additionally, a newly designed network will be analyzed. With the evaluation of three deep neural network structures extensive experiments on the performance of these methods in the task of cross-modal person re-identification are performed. Hence, it was possible to define a new state-of-the-art table for two datasets.

Second, an introduction of a new two-step deep neural network training scheme for cross-modal re-identification between depth and RGB was presented. This neural network exploits the relationship of the depth and RGB modality within cross-modal distillation. The cross-modal distillation network is considered the state-of-the-art for cross-modal person re-identification between depth and RGB. At the same time it is shown, that the newly presented architecture cannot directly applied to all cross-modal tasks as the performance in re-identification between infrared and RGB was not outperforming previous work.

Third, within the successful deployment of the cross-modal distillation network it was possible to contribute to a better understanding of the asymmetrical relationship

between depth and RGB modalities. It was shown, that features which can be extracted in the depth modality can up to a certain degree also be extracted in the RGB modality. This knowledge can be leveraged for future problems-solving approaches concerning the cross-modal relationship between depth and RGB.

3 Methods for cross-modal person re-identification

One of the contributions of this work is the deployment of deep neural network structures for the cross-modal person re-identification task in RGB and depth. Therefore, in the following three deep neural network structure for cross-modal person re-identification will be deployed on the task of cross-modal re-identification between depth and RGB. Two of these methods, the one-stream neural network and the zero-padding network are extracted from [83] as those methods were proven to be successful in cross-modal person re-identification between infrared and RGB. The third network, the cross-distillation neural network was developed in this work on the basis of the ideas of [77] and is considered another main contribution of this work.

3.1 One-stream neural network

The usage of a one-stream neural networks is the standard case in neural network training and deployment. Generally, inputs from a single modality are provided to the network and optimization is performed by hands of a standard loss, like softmax. To adjust for the cross-modal case, the input to the network is simply a mix of two modalities. This means that input images from the different modalities are given to the network in an equal manner. In the optimization process it is expected, that the network learns to embed the two modalities into a common feature space without any further guidance from the outside.

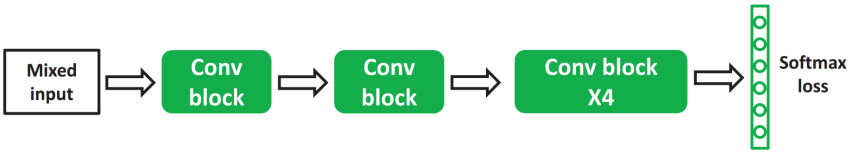


Figure 17: Exemplary structure of an one-stream neural network network [83].

An example scheme for an one-stream neural network can be seen in figure 17. The application of the one-stream network for cross-modal sensing is extracted from Wu et al. [83] where the network was used for sensing between RGB and infrared images. In this paper the neural network architecture will be differing slightly from [83]. Whilst Wu et al. [83] used a Resnet6 model for the evaluation of the one-stream network, here a Resnet18 and a Resnet50 structure will be analyzed. According to Wu et al. [83] the network will be optimized with softmax loss.

3.2 Zero-padding neural network

The basic idea of two-stream neural networks is to dedicate a separate part of the neural network to each of the modalities. The example of a two-stream network used in this work is the zero-padding neural network. This network structure was presented in chapter 2.3.3. The idea is to bring the images from the two sensing modalities to one channel. Hence, RGB is brought to grayscale, whilst the depth modality remains in its single

stream. Afterwards these one-channel images are combined with an empty or zero-padded channel. The channels are combined in a way, that each modality has its own separate channel. The approach is visualized in figure 14. Within this approach the network is supposed to have guidance on modality-specific nodes and shared nodes as can be seen in figure 18.

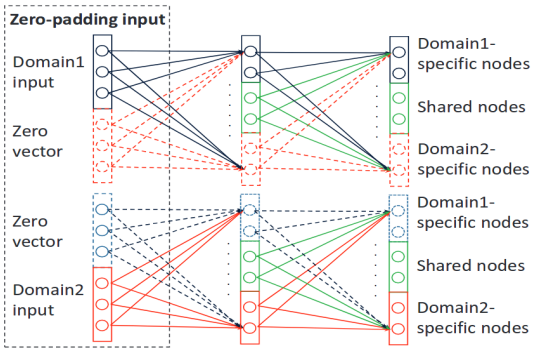


Figure 18: Zero-padding network. Visualization of domain-specific and shared nodes [83].

In the evolution of this work a second two-stream neural network structure was investigated. Unfortunately, it turned out, that the implementation as provided by the authors was not giving the expected results. More information on this two-stream structure can be found in appendix A.

3.3 A cross-modal distillation network

This subsection introduces our novel cross-modal approach. The major difference to the approaches presented in the previous subsection is that the tasks T^s from the source modality and the task T^t from the target modality are approached in a sequential manner, rather than in parallel. Therefore, the training of the task in the source modality is separated from the training of the task in the target modality. The cross-distillation scheme to transfer the supervision from one modality to the other modality is adapted from the work by Gupta et al. [77] (see section 2.3.1). To make use of the cross-distillation, the training of the network is divided into two steps, as it is visualized in figure 19, which will be explained in detail next.

3.3.1 Step I – Training of the baseline network

In step I of the training of the cross-modal distillation network, a neural network F is trained for sensing in a first modality, as presented in section 2.2.2. Therefore, we make use of a combination of feature extractors and losses as presented in chapter 2.2. The feature extractors Resnet18 and Resnet50 as well as softmax loss and triplet loss will be used to optimize networks for the baseline of the cross-modal distillation network in this work.

The network is optimized by hands of an early-stopping criteria based on the mAP in the validation set. Afterwards, the network is frozen as F_{fr}, with corresponding weights $W_{F,fr.}$.

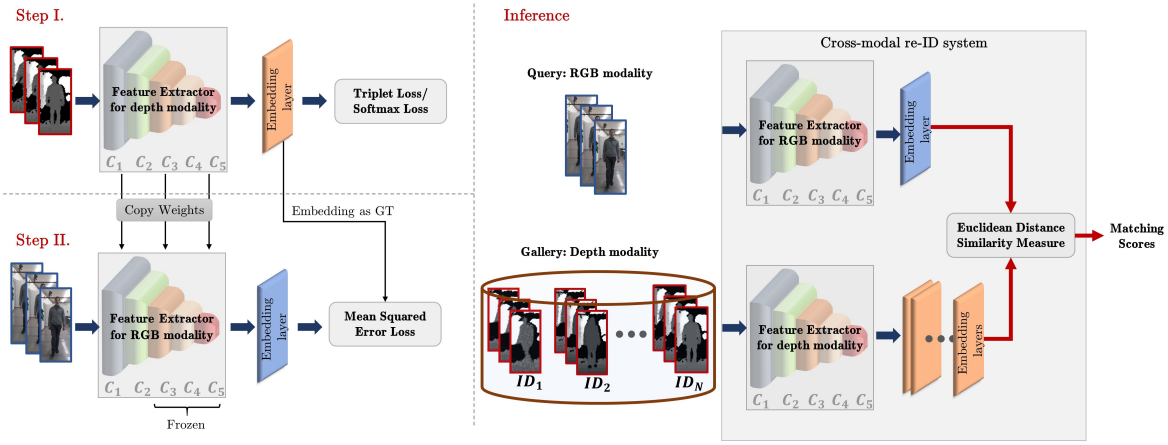


Figure 19: Two step training scheme and inference for the proposed cross-distillation network. Step I involves training of a CNN for single-modal re-identification. In step II, the knowledge from the first modality is transferred to the second modality. During inference, query and gallery images different modalities produce feature embeddings and matching scores for cross-modal re-identification. This figure is exemplary of a transfer from depth to RGB, and a inference with RGB as query and depth as gallery. The modalities can be interchanged in both cases.

3.3.2 Step II – Cross-Distillation

The obtained neural network feature extractor for the first modality is deployed as the baseline network for the training of a feature extractor for the second modality. For the second training step, a network with the same architecture as the corresponding network in step I is initialized.

Similarly to Gupta et al. [77], the weights of the converged model from step I, $W_{F,fr.}$, are copied to network G which is dedicated to the second modality. Additionally, the weights of the network are frozen from a mid-level convolutional layer up to the final feature embedding.

This retains the high-level mapping from the first network, which was successfully trained in the source modality, to the target modality.

At the same time, the target embedding can still learn meaningful low-level features for the task in the target modality.

For the actual transfer of knowledge we make use of paired images X_{m1} from modality 1 and X_{m2} from modality 2. The aim is to optimize G in such a way that the embeddings of images from the second modality X_{m2} with label y are close to the embeddings of images from the first modality X_{m1} with label y. This is realized by exploiting image pairs $x_{m1,i}$ and $x_{m2,i}$ from the two modalities, which are considered coupled as they are taken at the exactly same time step.

Hence, the embedding of $x_{m1,i}$ is obtained with a forward propagation through the frozen network $F_{m1,fr.}$ and is taken as the groundtruth for the embedding of $x_{m2,i}$ with the, at this stage, trainable network G. Since during inference mode the embeddings will be compared based on Euclidean distance, we aim to minimize this metric between the two embeddings. Hence, we make use of the mean squared error (MSE) loss between the

Algorithm 1 Cross-Distillation Method

Input: Input Train Data with paired images, X_{m1}, X_{m2}

STEP I:
$j = 0$
$mAP_{val,best} = 0$
Initialize network F with parameters W_F using a pre-trained CNN
while $(j < MAXEPOCH)$ **do**
    Perform training of F, train (X_{m1},W_F) using the loss functions in equations (2) or (3) or (5).
    if $mAP_{val,j} > mAP_{val,best}$ **then**
        save W_F as $W_{F,best}$
    end if
    $j = j + 1$
end while

STEP II:
$j = 0$
$L_{val,best} = \infty$
Load $W_{F,best}$ into F and freeze to $F_{fr.}$
Initialize weights W_G of network G with weights $W_{F,best}$
Freeze mid- to high-level weights of W_G
while $(j < MAXEPOCH)$ **do**
    Perform training of G_{m2}, train (X_{m2}, W_G) using loss function 6 and $F_{fr.}(X_{m1})$ as groundtruth
    if $L_{val,j} < L_{val,best}$ **then**
        save W_G as $W_{G,best}$
    end if
    $j = j + 1$
end while
Load $W_{G,best}$ into G and freeze to $G_{fr.}$
Output: Models $F_{fr.}$ and $G_{fr.}$

embeddings of paired images $F_{fr.}(x_{m1,i})$ and $G(x_{m2,i})$ which is defined as

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|F_{fr.}(x_{m1,i}) - G(x_{m2,i})\|^2, \tag{6}$$

where N is the batch size in training stage. The weights W_G of network G are optimized based on this loss function and trained until convergence. Early-stopping criteria for the training of this network is the loss in the validation set. The whole training procedure is formalized in algorithm 1.

3.3.3 Inference

In inference mode, the two resulting neural networks $F_{fr.}$ and $G_{fr.}$ are evaluated in the corresponding modalities to obtain feature embeddings for input images. Similarity between the feature representations of RGB and depth images is measured using Euclidean

distance. For each query image, each gallery image is therefore ranked according to the similarity between embeddings in Euclidean space.

4 Datasets, experimental methodology and experimental details

This work is targeting the previously almost untouched challenge of cross-modal person re-identification between RGB and depth. Therefore, it is necessary to have a detailed look at, on the one hand, evaluation methodology and on the other hand available datasets for the task. In the following, these two topics will be analyzed in detail. Additionally to that a detailed view on the experimental implementation which was used for this work will be taken.

4.1 Datasets

In the following, the datasets which are most relevant for this work will be discussed. Thereby, a discussion of splits in training, validation and test sets will be made. Additionally, a short glimpse will be taken at other datasets and it will be discussed why those were not considered for the validation of the results. As discussed in chapter 2.4 the focus is not be solely on re-identification between RGB and depth modality combinations, but also a common dataset for re-identification between infrared and the RGB modality will be considered to be able to discuss generalization capabilities of the presented techniques.

4.1.1 BIWI RGBD-ID Dataset

The BIWI RGBD-ID dataset was published by the Intelligent Autonomous Systems Laboratory (IAS-Lab) of the University of Padua in 2013. The dataset was developed to target long-term people re-identification from RGB-D cameras [87].

The dataset was recorded with a Microsoft Kinect and contains a total of 50 different people which are divided into 50 training and 56 testing sequences. The 56 testing sequences consist of 28 persons with two sequences each. The testing sequences were recorded on a different day and in a different location and, therefore, contain variations, like different clothing. In the training sequences the persons perform a certain routine of motions such as head movements and walking to the camera. The testing sequences consist of "Still", which contains no movement and "Walking", where a walk in direction camera is performed. The amount of images contained in the different modalities can be seen in table 1. Exemplary images from the BIWI-RGBD-ID dataset can be seen in figure 20 [87].

The BIWI RGBD-ID dataset was used as validation for the cross-modal re-identification methods presented in [5], which is the only comparison paper for cross-modal re-id in RGB-depth. A more detailed discussion of the results in this paper will be made in chapter 5.3.

When analyzing the literature, in which the BIWI RGBD-ID dataset is used, it gets visible, that the originally presented split in training and testing set of the authors was neglected in most of the follow-up papers. Wu et al. [78] used only the 22 persons with only one sequences and without appearance change for training, and the 28 persons with three sequences for testing. Liu et al. [88] made the assumption that the same person wears the same clothes. Therefore, they discarded the designated training instances and only used the still and walking sequences from the designated testing set. Zhuo et al. [5] take the same assumption and go one step further. They split the 28 persons

with different clothing and define them as different persons when they wear different clothes. In this manner they manage to obtain 78 persons instances (28x2 + 22). They divide these 78 persons randomly in 40 persons for training and 38 for testing. For the comparison in this work, the training set of 40 person will be divided into 32 training and 8 validation persons. All images are synchronized.

Models which are not based on hand-crafted features, like neural networks, are very dependent on training data which captures the same distribution as the test data. Therefore, it was logical, that the provided training-test split had to be broken up as no sequence of changing clothes was part of the training set. On the other hand, it contradicts the intention of the authors of the dataset to design a long-term oriented re-identification system, to simply combine the persons with different clothing as was done in Zhuo et al. [5]. On the other hand, in the relevant scenarios in surveillance and autonomous driving a clothing change is very unlikely. Additionally, Zhuo et al. [5] is the only comparable source for cross-modal re-identification (see chapter 2.3). Hence, in this work this split will be evaluated. The split by identities can be found in appendix B.1.

Figure 20: Example images from BIWI [87]. First and third image from the RGB modality. Second and fourth image from the depth modality. Images are coupled.

4.1.2 RobotPKU RGBD-ID dataset

Another dataset with similar characteristics to the BIWI dataset is the RobotPKU dataset. It was published by Liu et al. in 2017 [88].

The dataset consists of 90 persons. In the original paper no split in training, validation and test set was presented. For this work the division will be 40, 10, 40 for training, validation and test set, respectively. This follows the division of Liu et al. [88] the best as possible. They reported a division of training and test set in 50 persons each. The split and the corresponding amount of images can be seen in figure 1. Example images are shown in figure 21.

The RobotPKU dataset is a more challenging dataset than the BIWI RGBD-ID dataset for two reasons. Firstly, the depth images dataset are much more error-prone. For example, in figure 21 in the depth image on the very right the head of the person is not entirely captured by the depth device. Errors like this are much rarer in the BIWI dataset. Secondly, the images are not perfectly coupled like in BIWI and a small difference within several milliseconds between depth and RGB images are possible. Hence, the performance test on the RobotPKU dataset can be considered a robustness test for datasets which are well-performing on BIWI.

The split by identities which is used in this work can be found in appendix B.2.

Figure 21: Example images from RobotPKU dataset [88]. First and third image from the RGB modality. Second and fourth image from the depth modality.

4.1.3 SYSU RGB-IR Re-ID

As was discussed in chapter 1, it is necessary to take a look at different modality combinations to obtain a thorough overview over the capabilities of a method for cross-modal re-identification. For this reason the dataset SYSU RGB-IR Re-ID [83] will be considered.
The dataset combines the modalities visible light and infrared (IR) light. It consists of 491 identities from 6 cameras, giving in total 29,023 RGB images and 16,579 IR images (see also table 1). Camera 3 and 6 are capturing infrared, while the other 4 cameras capture RGB images. Camera 1,2 and 3 are placed indoor, while camera 2 and 3 are in the same room. The remaining cameras are capturing an outdoor environment. All together there are 496 identities in the dataset, of which 296 identities are used for training, 99 for validation and 96 for testing. The images are captured in bright (RGB) or dark (IR) environment and, hence, not synchronized. Example images can be seen in figure 22.

There are two scenarios given for testing: 1. The 'All-search' scenario, where all cameras are used. 2. The indoor scenario, where only cameras 1, 2 and 3 are relevant. In both cases the visible light images are used as gallery and the infrared images as probe.

Figure 22: Example images from SYSU RGB-IR Re-ID dataset. Top images from visible light modality, bottom images from infrared modality [83].

4.1.4 Other Datasets

Besides the presented datasets especially in the RGB-D domain some additional datasets were found. For the IIT RGB-D dataset [69] the problem was, that the depth data was given as a pointcloud and very few images per person were available. For this work,

Dataset	train				val				test				overall			
	M1 #ids	M1 #imgs	M2 #ids	M2 imgs	M1 #ids	M1 #imgs	M2 #ids	M2 imgs	M1 #ids	M1 #imgs	M2 #ids	M2 imgs	M1 #ids	M1 #imgs	M2 #ids	M2 imgs
BIWI	32	9245	8	9245	8	2097	8	2097	38	10696	38	10696	78	22038	78	22038
RobotPKU	40	7400	40	7400	10	1815	10	1815	40	7297	40	7297	90	16512	90	16512
SYSU-IR	296	20274	296	9929	99	1974	99	1980	96	6775	96	3803	491	29023	491	15712

Table 1: Overview over the datasets. For BIWI and RobotPKU: Modality 1 (M1) is RGB, Modality 2 (M2) is Depth. For SYSU-IR: Modality 1 (M1) is RGB, Modality 2 (M2) is infrared.

it was considered logical to focus on datasets which do not necessarily need additional preprocessing steps to feed the images to a neural network. Additionally, preliminary tests showed, that the images were too few for successfully training deep learning methods.

The TUM Gaid dataset [70] includes around 300 persons and is, therefore, theoretically very well suited for deep learning. Unfortunately, the depth images in the dataset are very small and preliminary tests showed, that no meaningful insights could be achieved on this dataset.

4.2 Measures of performance

The evaluation of the validation and test loss in re-identification tasks is more challenging, than in most other machine learning tasks as the persons contained in the validation and test set are not part of the training set. Therefore, it is not possible to directly measure the validation or test loss at a certain time step like in classical classification tasks. For example, for a classification task optimized with softmax loss, the loss can simply be measured by the difference between the obtained probability for a class and the groundtruth class label. In a re-identification task the groundtruth labels of validation and test set are not part of the classification layer as a preliminary embedding layer is used for comparison. Hence, the performance of a re-identification network has to be monitored by taking a look at final evaluation measures.

The most common evaluation measures in re-identification are the cumulative matching characteristics (CMC) and the mean average precision (mAP). These will be presented in the following. Beforehand two concepts have to be introduced, which are closely related. Additionally, a look at a technique for deconvolution of neural networks for visualizing the most activating regions for a network in an image will be discussed as this will be used for a qualitative analysis of the results.

4.2.1 Probe/Query vs. Gallery/Target set

In general in evaluation settings for re-identification a target set and a query set are differentiated. Both sets contain images of the same person instances which are extracted from validation or tests set. Generally, out of the target set a gallery set is constructed and from the query set the probe set is obtained. Despite this definition, the words "query" and "probe" set as well as "gallery" and "target" set are often used interchangeable.

The gallery set can be considered the comparison set for the probe set. Usually, the output of an algorithm is a similarity measure which defines the distance between all images from the target set to the query set [71]. From this similarity matrix the measures which will be presented in chapters 4.2.3 and 4.2.4 are calculated.

4.2.2 Single-gallery shot vs. multi-gallery shot

While the definition of the query set is constant within the literature, the construction of the gallery set can be done in two manners.

In a single-gallery shot setting the gallery consists of one, normally randomly chosen image from each object in the gallery. Therefore, only one image of each person in the gallery set is used for the evaluation of the measures. Afterwards, the distance of each image in the probe is calculated (one-to-many comparison) to each image in the gallery and the statistics CMC and/or mAP (see chapters 4.2.4 and 4.2.3) are calculated.

In a multi-gallery shot setting more than one image of an individual object can be part of the gallery set. This leads to a higher diversity in the gallery set. An usual effect of having more than one image in the gallery set is, that higher accuracies (especially for CMC) can be reached. Intuitively, this can be explained by the fact, that it is more likely that an 'easy' image of the same person is part of the gallery set.

4.2.3 Cumulative Matching Characteristics (CMC)

As re-identification is an inherently difficult task, it mostly does not make sense to only consider if the top match between query image and gallery images is correct. Therefore, the cumulative matching characteristics curve (CMC) describes, if the correct match is among the first k matches. CMC basically shows the probability for an image in the probe set to find a correct match among the first k most probable matches in the gallery. It is possible to indicate the performance with the area under the CMC curve. Nevertheless, in more recent literature it is more common to report several of the top k ranks instead of one value for CMC. Popular values for k are 1, 5 and 10 (R1, R5 and R10) [34, 83, 78]. CMC values can be very biased by multi-shot settings, as they are only reporting the first match of a probe image. Assuming a decently performing model, the more images of the same person are in the gallery, the higher is the probability that the first match of a probe image is from the same class.

4.2.4 Mean Average Precision (mAP)

Before the publication of Market-1501 [34], it was common in person re-identification to evaluate the performance of algorithms solely based on CMC indicators. Zheng et al. [34] argumented the need for another statistic with the scheme which can be seen in figure 23. In the example cases (a), (b) and (c) the CMC rank 1 accuracy is always 1, because the first image is a match in all cases. Nevertheless, there is an apparent difference especially between (b) and (c), where two images of the same person are in the gallery (multi-gallery shot). For a fair comparison, it is necessary to differentiate if the second gallery item is detected on position two, like in (b) or position five, like in (c). In this case, "recall", which is defined as the proportion of true positives not identified, is not considered in CMC and average precision is better applicable for measurement [34].

The evaluation of machine learning techniques with a precision-recall curve is a popular instrument. For the calculation it is necessary to identify true positives (tp), false negatives (fn) and false positives (fp). Precision is defined as $\frac{tp}{tp+fp}$, while recall is defined as $\frac{tp}{tp+fn}$. The statistic is often displayed as a step function for different thresholds with precision on the y-axis and recall on the x-axis. The average precision (AP) is defined as the area under the precision-recall curve. In re-identification the threshold is defined as the distance of the probe image to gallery images which are accepted as the same person.

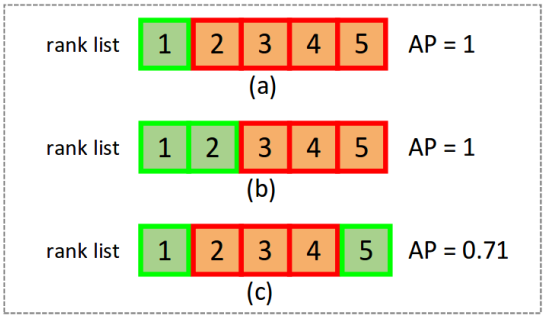


Figure 23: Problematic nature of using solely CMC curve for measurements. While CMC is 1 for all cases, AP additionally captures recall (in (c) only 0.71 accuracy). Green is same person image, red is other person. Source: [34].

Therefore, the average precision has to be calculated separately for each probe image. Finally, the mean of the average precision of all probe images are reported as the mean average precision (mAP). MAP is less influenced by single- versus multi-gallery shot settings than CMC, as it inherently evaluates the performance over all positive samples.

4.2.5 Deconvolution of neural networks

A visualization of the activations of a neural network can provide insights into the learned patterns of a network. Therefore, in this work one method for activation visualization will be used. In 2014, Springenberg et al. [82] presented an approach to visualize the concepts learned by higher neural network layers in a simple and efficient way. The idea is to invert the data flow of a convolutional neural network by move from neuron activations on a specific higher-level layer down to an input image. This process is called deconvolution. In the approach Springenberg et al. [82] presented, the idea is to use an image as well as its groundtruth to get a reasonable reconstruction of the image. The deconvolution step itself is analogous to a backward pass through the network. The main difference to a typical backward pass is, that when propagating through a non-linearity, like a rectified linear unit layer, the gradients are computed based on only the top gradient signal. The authors call this guided backpropagation, because it adds an additional guidance from higher layers. The idea behind it is to diminish the influence of negative gradients, which decrease the activation of units, which are meant to be visualized.

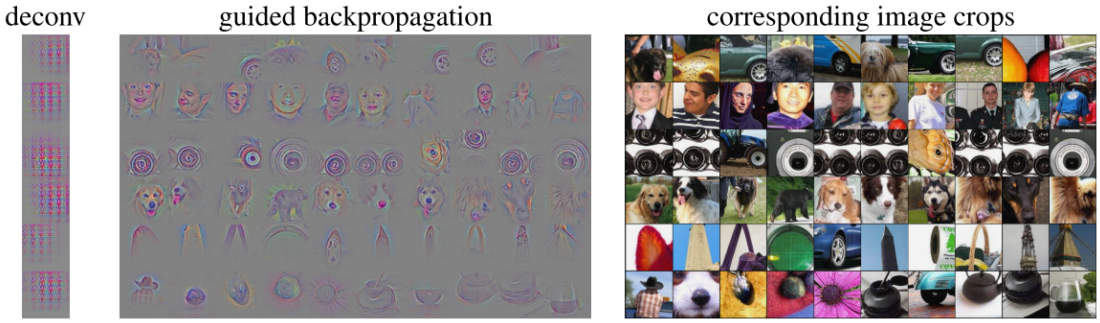


Figure 24: Example deconvolution results with guided backpropagation. Source: [82].

In figure 24 some example results of guided backpropagation are visualized. It gets visible, which parts of the images are activating the network and lead to the final classification decision, like e.g. the dog noses for the classification as a dog. On the very left, the image shows the results, when the guidance is taking out of the deconvolution algorithm. The obtained images are much less descriptive than the ones for guided backpropagation.

4.3 Experimental Details

After having discussed the main datasets and measures of performance for the experimental chapter, it is possible to take a look at details of the experimental implementations. This will be done in the following and is split into details on evaluation and details on the training procedures.

4.3.1 Details on Evaluation

For this work, the datasets for testing the methods are split into training, validation and test set. Even though this is considered a good practice in machine learning research [47], many methods in the area of person re-identification are ignoring this separation and only use a training and a test set [5]. Therefore, those methods are highly prone to overfit to the test data. For the BIWI RGBD-ID dataset and the RobotPKU dataset no splits of the dataset by the original authors were provided and the dataset had to be split by the author of this paper (for splits see Appendix B). To get meaningful performance indicators for these datasets a 3-fold cross-validation procedure was followed for all tests. This means, that three times a different validation set was extracted from the design set. For the SYSU-IR dataset a split in training, validation and test set was provided by the original authors. Hence, no further cross-validation was performed.

Especially for the datasets with a high amount of images for each individual, it was necessary to restrict the amount of images which are taken into account for the evaluations to get a reasonable trade-off between training time and evaluation time. Following the procedure used in [83] for each evaluation of mAP, a maximum of 50 random images of each person in the probe. It was shown that the difference to a scenario that considers all images is minimal. For evaluation of Rank1, Rank5 and Rank10 the same excerpt of images is taken into account. In general for person re-identification tasks, images from the same camera are taken out of the evaluation, following the evaluation protocol in CUHK03. For the BIWI RGBD-ID and the RobotPKU dataset the camera constraint had to be relaxed, as for most identities only one camera view is available. The same holds for the validation set of SYSU RGBD-ID.

Additionally, for each evaluation only one of the images of one instance in the gallery is used to obtain a single-shot setting. In an evaluation different images are taken for the gallery set and the evaluation is repeated 10 times.

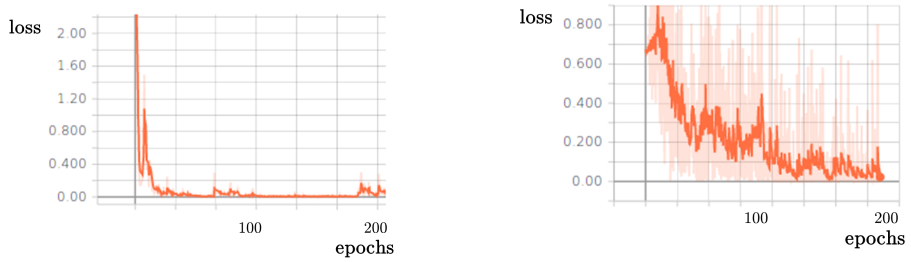
To compare the performance the rank 1 accuracy on the test set, rank 5 accuracy on the test set, rank 10 accuracy on the test set and mAP on the test set will be reported.

For some networks in the following chapters further visualizations of the networks are available in the appendix.

4.3.2 Details on Training procedures

In general, in the following the neural networks are optimized with an an early stopping criteria. For all datasets the early stopping criteria is the mean average precision (mAP, compare chapter 4.2.4) of the validation set. The neural networks taken into account for testing are Resnet18 and Resnet50 as explained in chapter 2.2.2. Those networks have shown very good performance in re-identification tasks and are, therefore, taken as the baseline for this work. The networks are optimized with softmax and triplet loss. The neural networks trained with softmax were optimized with stochastic gradient descent with Nesterov momentum [89]. Those trained with triplet loss were optimized with the ADAM optimizer [90]. For the next chapter all networks were designed to obtain a feature size of 128 as suggested in the literature [45]. A dropout rate of 0.5 was deployed. The margin for triplet loss (see formula 5) was set to 0.5.

The training of the neural networks for re-identification in the different domains are classical neural network optimizations. Therefore, a training loss is optimized and ideally converges to zero. An example loss for a successful optimization with triplet loss and softmax loss can be seen in figure 25. Even though in an optimization for triplet loss more variation in the graph is visible, both curves converge to zero.



(a) Loss curve in successful training with softmax loss.

(b) Loss curve in successful training with triplet loss. Curve is smoothed out to make trend more visible.

Figure 25: Loss curves for successful trainings with triplet and softmax loss.

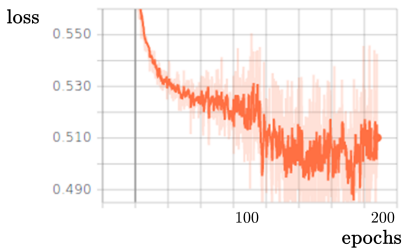


Figure 26: Loss curves for unsuccessful training of triplet loss

The main reason, why the loss curves have to be discussed in detail, is that there is a certain complexity in training neural networks with triplet loss. In some cases the training can lead to a training loss as it can be seen in figure 26. Here, the training loss is not converging to zero but to 0.5. To explain this behaviour a view has to be taken on formula 5 which is the definition of triplet loss. It consists of three parts. Firstly, the distance of the embeddings of the anchor and positive. Secondly, the distance of the embeddings of the anchor and negative. This part is subtracted from the first part and,

thirdly, a margin α is added. As defined in the introduction of this chapter, α is set to 0.5 for the presented trainings. Clearly, in a successful training the distance between the two former parts of the formula is bigger than the margin 0.5 and, therefore, a loss of 0 is achievable. But, and this is what happened in the training shown in figure 26, if the task is difficult to solve, the optimization can lead to a network which is embedding all images in the same point and, hence, obtains a distance of 0 for all images. In this case the loss of the network simply converges to α. This is dangerous because of two reasons. Firstly, the performance in the validation set of the networks in re-identification tasks is measured with mAP because it can not be measured with a loss directly connected to the training loss (see chapter 4.2). Hence, the problem can only be identified in monitoring the training loss. Secondly, the standard evaluation functions for mAP and CMC are vulnerable to embeddings of all images to zero, as they simply look for the least distance and do not include a sanity check.

There are some potential solutions to the problem, like more careful initialization and training of the networks. As simple variations did not lead to a solution for the considered cases and the goal of this work is not to optimize training procedures with triplet loss, the complexity will be accepted and if a network training is not possible it will be indicated as n/a in the evaluation.

5 Experimental Results

In the following sections the experimental results of this work will be presented in two subsections. The first subsection presents the results for optimizing deep neural networks in single modalities. This analysis functions as a comparison baseline for the rest of this work, and at the same time corresponds to step I of the cross-modal distillation network (section 3.3). In the second subsection the results for the three neural network techniques for cross-modal sensing will be presented.

5.1 Optimization in single-modal re-identification

In section 2.2 several methods for optimizing a neural network for person re-identification in a *single* modality were given. To get a guideline for this paper in the following several neural network architectures and optimization techniques will be compared in the single-modal task. This analysis will be used as a comparison baseline for the cross-modal methods and acts as step I for the cross-distillation network. The results for the neural networks presented in this section are explicitly not optimized for cross-modal sensing and solely optimized for re-identification in one single modality.

5.1.1 BIWI RGBD-ID dataset

The BIWI RGBD-ID dataset was presented in section 4.1.1. It consists of depth and RGB images. For the tests which are presented in table 2 the split of 32 individuals for training, 8 individuals for validation and 38 for testing was taken.

Table 2: Average test set accuracy of different deep neural network architectures in the single-modal task for the BIWI dataset.

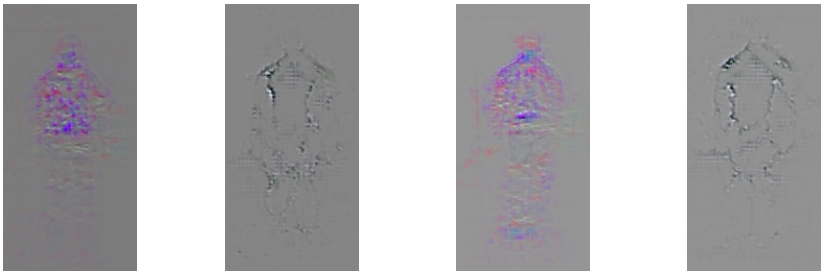
Modality	**Feature Extractor**	**Loss**	**R1 (%)**	**R5 (%)**	**R10 (%)**	**mAP (%)**
RGB	Resnet18	Triplet	93.68 ± 0.76	99.65 ± 0.35	99.96 ± 0.04	94.77 ± 0.83
		Softmax	93.32 ± 1.83	99.67 ± 0.24	99.93 ± 0.09	94.46 ± 1.55
	Resnet50	Triplet	92.14 ± 1.86	99.71 ± 0.24	99.95 ± 0.08	93.44 ± 1.46
		Softmax	**94.75 ± 0.74**	**99.75 ± 0.19**	**99.96 ± 0.03**	**95.68 ± 0.60**
Depth	Resnet18	**Triplet**	**61.28 ± 2.49**	**93.85 ± 1.05**	**99.44 ± 0.18**	**62.71 ± 2.37**
		Softmax	57.09 ± 0.79	88.96 ± 0.15	96.95 ± 0.20	58.38 ± 1.07
	Resnet50	Triplet	54.23 ± 1.75	91.48 ± 0.56	99.15 ± 0.18	55.31 ± 1.71
		Softmax	59.84 ± 0.66	90.54 ± 0.81	97.80 ± 0.19	61.44 ± 0.54

In RGB the performance of the classifiers are very good. All trained models obtained an mAP of 93% or higher on the test set. The best model is the Resnet50 network optimized with softmax loss with an average mAP of 95.68%. Resnet18 with softmax loss obtains a mAP of 94.46% and, hence, is competitive to the deeper version. The networks trained with softmax loss obtain an average mAP of 94.77% for Resnet18 and 93.44% for Resnet50. Hence, in this case the shallower network is better suited for the task as the deeper one.

The average mAPs in the depth domain are much lower than in RGB. Here, the best performing model is Resnet18 trained with triplet loss which achieved an average mAP of 62.71%. Again, the deeper Resnet50 model performed worse for the optimization with triplet loss. For optimization with softmax loss Resnet50 outperformed Resnet18 by around 3%. The average mAP of Resnet50 was 61.44%. Hence, the best models for re-identification in pure depth were around 30% worse in average mAP in comparison to

the best models in RGB. This indicates, that the re-identification task in depth is more difficult to solve in comparison to the re-identification in visible light. In figure 43 and 44 examples for query-gallery results are shown. Analyzing the two tasks visually it gets clear, that also for humans the task in depth is much more difficult to solve than the task in RGB.

To understand how the neural network solved the re-identification tasks, a possibility is to analyze which parts of the images activated the neural networks the most. In section 4.2.5 a method to obtain such gradient images via guided deconvolution was presented. In figure 27 the gradient images of the images from figure 20 are shown. All images are calculated with Resnet18 trained with softmax loss for the corresponding modalitites.



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 27: Single-modality networks: Gradient images for BIWI with Resnet18 and softmax loss.

It is shwon, that a big difference between the activations of a RGB image and a depth image exist. In RGB images ((a) and (c) in figure 27) bigger regions are activating the network. Those regions can be identified as the head region and the torso of the person. It is quite apparent, that the overall appearance of the torso, including colors is evaluated. This is reasonable as the torso is fully visible in almost all images, and mostly not changing its appearance that much in frontal (a) and dorsal images (c). The high variability in the visibility of the legs of a person when walking, leads to the fact, that the image is lowly activated by this part.

In images (b) and (c) the activations of the networks trained on the depth images are visible. The activation functions are much differently built than for the RGB images. To extract features for classifying the depth images, the network is dependent on boundary structures of the body. The images suggest that two structures are used. First, the overall body shape of torso, arms and upper legs. Again the lower part of the legs is mostly neglected. Most probable, because this part is too variable. Second, it seems that the network is most activated by the structure of the torso. A reason for this can be, that the torso contains most of the recurring describing features for a depth image. Almost no coherence between the activation maps of depth and RGB images can be identified.

5.1.2 RobotPKU dataset

The Robot PKU dataset is similar to the BIWI dataset and consists of RGB images and depth images. As explained in section 4.1.2 the depth images are more noisy and, hence, solving the re-identification task in RobotPKU is more complex. Exemplary images for RGB and depth are shown in figure 21.

For the RGB modality an average mAP of up to 91.91% is reached with a Resnet18

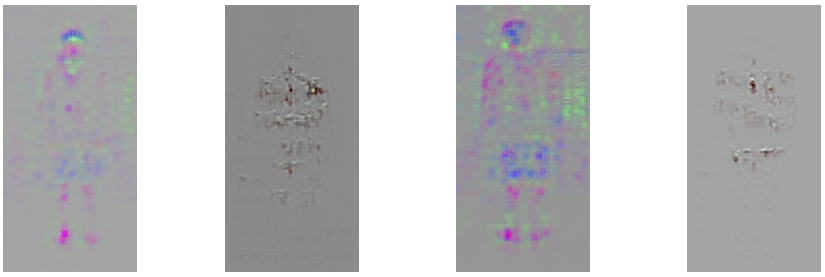
trained with triplet loss. The deeper Resnet50 performs worse and obtains an average mAP of 90.63%. Both networks trained with triplet loss outperform the corresponding networks trained with softmax loss. Here, Resnet18 and Resnet50 obtain an average mAP of 86.86% and 87.11%. Similarly to the BIWI dataset, the performance of all networks in the RGB modality is high.

Networks optimized to perform in the depth modality are much more difficult to train. It was not possible to successfully train a Resnet18 and Resnet50 with a triplet loss. For a more thorough explanation on why this was not possible please refer to section 4.3.2. The same networks trained with softmax loss obtained an average mAP of 38.65% for Resnet18 and 44.03% for Resnet50. These values are around 40% lower than for the RGB modality. Hence, a significant gap between the difficulty of the separate re-identification task in RGB and in depth exist. Also in comparison to the performance in the depth modality in BIWI (table 2) a gap of 20% exists, while the amount of instances in the test set are almost same. This indicates that the task in the depth modality for RobotPKU is more difficult than for BIWI. Nevertheless, comparing the results to random guessing, which lies at 2% for Rank1 accuracy, the results are still acceptable.

Table 3: Average test set accuracy of the different deep neural network architectures in the single-modal task for the RobotPKU dataset.

Modality	**Feature Extractor**	**Loss**	**R1 (%)**	**R5 (%)**	**R10 (%)**	**mAP (%)**
RGB	Resnet18	Triplet	**90.53 ± 0.65**	**99.30 ± 0.17**	**99.46 ± 0.10**	**91.91 ± 0.64**
		Softmax	84.73 ± 0.47	98.00 ± 0.12	99.24 ± 0.14	86.86 ± 0.46
	Resnet50	Triplet	89.04 ± 3.91	99.17 ± 0.33	99.46 ± 0.10	90.63 ± 3.41
		Softmax	84.52 ± 0.24	97.91 ± 0.35	99.12 ± 0.23	87.11 ± 0.22
Depth	Resnet18	Triplet	n/a	n/a	n/a	n/a
		Softmax	39.17 ± 0.34	69.85 ± 0.63	82.58 ± 0.35	38.65 ± 0.44
	Resnet50	Triplet	n/a	n/a	n/a	n/a
		Softmax	**44.50 ± 1.02**	**75.83 ± 1.29**	**87.56 ± 0.87**	**44.50 ± 1.02**

To get more insights into the results, again a look at the gradient images will be taken. The images can be seen in figure 28 and correspond to the original images in figure 21. All gradient images are calculated with the Resnet18 networks trained with softmax in the corresponding modality. For RGB the networks are mainly activated by head and



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 28: Single-modality networks: Gradient images for RobotPKU with Resnet18 and softmax loss.

legs of the persons (see (a) and (c)). The torso seems to not play a very big role in the evaluation of the frontal image. Still for the dorsal image, a higher activation for the torso and the arms are found.

Again, the activations for the network in the depth modality are much different from

RGB (images (b) and (c)). Even though the very much cluttered structures in the image are difficult to interpret, it is possible to derive one main finding. The activations are mainly found in the region of the torso of the person. This is a big difference to the activations in the RGB modality. Therefore, this can be interpreted as a first indicator, that the inherently learned attributes to describe the persons of the two networks are not modelling the same features.

Further visualizations for the RobotPKU dataset and can be found in the appendix. In figure 47 and 48 exemplary query-gallery results are found. Looking at figure 48 can give a hint why the performance in the depth modality is generally low. For example in the second row the query image almost contains no information as many outer parts of the body are not captured by the depth-capturing device. This problem of the RobotPKU dataset was also described in section 4.1.2. Figure 47 gives insights in why for the RobotPKU dataset the torso is not as describing as in the BIWI dataset for the RGB images. The reason is simply that several persons wear very similar clothes. Therefore a more full view of the person is taken into account by the neural network comparison to the BIWI dataset.

5.1.3 SYSU RGB-IR dataset

The SYSU RGB-IR dataset consists of two modalities. The first modality is infrared images and the second modality is visible light images. The dataset is discussed to analyze the generalization of the presented methods for general cross-modal person re-identification. As for the SYSU dataset a fully defined split in training, validation and test set exists no standard deviation is reported for the accuracies. Example images can be seen in figure 22. In table 4 the results of the networks optimized for the single modalities are shown.

Table 4: Average test set accuracy of the different deep neural network architectures in the single-modal task for the SYSU-IR dataset.

Modality	**Feature Extractor**	**Loss**	**R1 (%)**	**R5 (%)**	**R10 (%)**	**mAP (%)**
RGB	Resnet18	Triplet	74.00	93.98	97.52	74.85
		Softmax	67.28	87.40	92.07	68.38
	Resnet50	Triplet	n/a	n/a	n/a	n/a
		Softmax	75.06	91.35	94.76	76.09
Infrared	Resnet18	Triplet	61.45	88.92	94.52	62.11
		Softmax	62.03	87.08	93.28	63.24
	Resnet50	Triplet	n/a	n/a	n/a	n/a
		Softmax	68.58	91.16	96.09	69.91

In the RGB modality the best performing model is a Resnet50 trained with softmax loss. It obtained an average mAP of 76.09%. The shallower Resnet18 reached an mAP of 7% less. With triplet loss only Resnet18 was trained successfully, it obtained an mAP of 74.00%.

For the infrared modality Resnet18 and Resnet50 trained with softmax loss perform with 63.24% and 69.91%, respectively. Again, only a Resnet18 was trained successfully with triplet loss. It obtained an accuracy of 62.03%. Similarly, to the depth modalities in the other datasets, the infrared modality is more difficult to classify than the RGB modality. In the appendix in figure 39 and 40 exemplary gallery-query results are shown for Resnet18 trained in RGB and in infrared with softmax, respectively. The images show

the high complexity of the tasks.

An explanation of the lower performance in the RGB dataset in comparison to the datasets which were investigated earlier is the higher number of individuals of the test set in the SYSU dataset. With around 70% mAP the networks are still performing well.

5.1.4 Discussion

It was shown, that for all relevant datasets in both modalitites for the single-modal case networks could be trained which obtained decent to good results. The datasets have in common, that for the RGB modality a high re-identification performance in terms of test mAP can be achieved. For all examined datasets, the performance in the RGB modality was higher than for the corresponding modality. For depth (BIWI and RobotPKU) the difference was more significant than for infrared (SYSU).

It is important to understand, that the results of this section are an indicator for possible performances of the methods for cross-modal sensing in the following chapters. For example, Resnet50 trained with softmax in RobotPKU obtains a rank 1 accuracy of 84.73% for RGB and 39.17% for depth, respectively. Hence, for cross-modal sensing the lower of the two values, in this case 39.17%, can most likely be considered an upper bound for the Rank 1 accuracy in cross-modal re-identification. This is logical, as individuals which are not re-identified in the same modality correctly can most likely not be re-identified when compared to an object from another modality. Hence, a hypothesis which can be made after this section is that sensing between modalities cannot be superior to sensing within the more challenging of the two modalities.

A look at the gradient images for the depth and RGB modalities gave first insights in the inherent modelling of the networks. It was found, that networks trained in the RGB modality are not automatically modelling structural features of the human body. The most activated regions are recurring attributes like colors or salient features, like shoes or heads. On the other hand, networks trained in the depth modality are only capable of sensing structural shape features. Overall, it got visible, that the learned features are very dataset dependent. Especially for the RGB modality the results for BIWI and RobotPKU were very different. The network trained with the BIWI dataset was very activated by the colors of the torso, while the network trained in RobotPKU was much more activated by heads and legs.

This analysis gives a first important hint for the next sections. It was shown, that separately trained networks for the domains use a very different base for the classification of the images. Therefore, it can be problematic to map the extracted features into a common feature space.

Another interesting finding was found for the interaction of a shallow and deep Resnet architecture with softmax and triplet loss. For all datasets, a network optimized with triplet loss was obtaining better accuracy when it was shallower (Resnet18). Networks optimized with softmax loss obtained better accuracies with a deeper architecture (Resnet50).

5.2 Optimization in cross-modal re-identification

In section 5.1 the capabilities of neural networks for re-identification in single modality were analyzed. It was shown, that in the single modal sensing task decent accuracies can be achieved. In this section the three neural network methods which were presented in section 3 will be evaluated on the datasets. After each subsection a short discussion on the results for the specific network architectures will take place.

In the following a lot of references to the preceding section 5.1 will be made. These methods will be named as *single-modal networks*. As the networks which are presented in the following are also capable of sensing in a single modality (indicated as single-modal task) references to section 5.1 with *single-modal networks* will be written in italic letters to avoid confusion.

5.2.1 One-stream neural network

The most straightforward method to allow cross-modal re-identification in deep neural networks are one-stream neural networks. The class of networks was presented in chapter 3.1. In this chapter the performance of one-stream networks on the presented datasets in chapter 1 will be reported. For evaluating the performance of the one-stream neural networks, several performance indicators will be evaluated. The networks sensing in the cross-modal task are inherently capable of sensing in single modalities as well. Hence, the performance in the test set in both cross-modal tasks (changing query and gallery), as well as in the individual modalities separately are reported. All these performance indicators are then used to discuss the behavior of the one-stream network.

In the literature the one-stream network is optimized by hands of a softmax loss. This will be followed in this work and similarly to the preceding section the feature extraction architecture will be varied between Resnet18 and Resnet50.

5.2.1.1 BIWI RGBD-ID dataset

Table 5 shows the results on the test set for the BIWI RGBD-ID dataset. The result for the cross-modal tasks can be seen in the first two columns for each architecture. For Resnet18 an average mAP of 14.55% for RGB as query and depth as gallery are obtained. For depth as query and RGB as gallery 20.09% in average mAP are obtained. The same measures for Resnet50 are 16.86% for RGB as query and depth as gallery and 23.75% for depth as query and RGB as gallery. Hence, for the cross-modal task the one-stream method profited from a deeper architecture within the BIWI dataset.

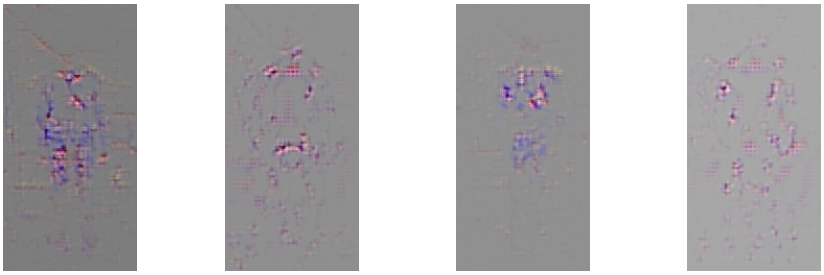
The bottom two lines (indicated as Q:RGB, G:RGB and Q:Depth, G:Depth) indicate the performance in the single-modal task for the networks. For Resnet18 in pure RGB an average mAP of 88.15% is obtained and for sensing in pure depth an average mAP of 54.23%. In the *single-modal networks* with the same architecture (see table 2) an average mAP of *94.46%* and *58.38%* was achieved. Hence, the performance of the one-stream network in the single-modal task is only slightly deteriorated by a few percentage points. For Resnet50 pure sensing in the one-stream network led to a performance of 90.16% and 59.06% for RGB and depth, respectively. Again this is only slightly inferior to the *95.68%* and *61.44%* in the *single-modal network*.

Again a look at the gradient images gives more insights into the activations of the network. In figure 29 the gradients are visualized. The corresponding original images are shown in figure 20 in chapter 4.1.1. Comparing the gradient images to the gradient

Architect.	Datas.	Loss	Feature Extractor	Inference modality	R1	R5	R10	mAP
One-stream network	BIWI	Softmax	Resnet18	Q:RGB, G: Depth	13.27% ± 2.54%	47.66% ± 4.37%	72.71% ± 3.09%	14.55% ± 1.99%
				Q:Depth, G: RGB	15.85% ± 1.77%	51.32% ± 1.87%	75.30% ± 2.61%	20.09% ± 1.32%
				Q:RGB, G: RGB	86.24% ± 2.51%	98.10% ± 0.65%	99.52% ± 0.13%	88.15% ± 2.07%
				Q:Depth, G: Depth	52.78% ± 1.76%	86.84% ± 0.99%	96.10% ± 0.38%	54.23% ± 2.07%
			Resnet50	Q:RGB, G: Depth	15.68% ± 0.77%	50.29% ± 1.18%	75.65% ± 0.46%	16.86% ± 0.87%
				Q:Depth, G: RGB	19.82% ± 0.33%	55.74% ± 0.83%	78.92% ± 1.07%	23.75% ± 0.30%
				Q:RGB, G: RGB	88.60% ± 1.67%	98.37% ± 0.13%	99.53% ± 0.06%	90.16% ± 1.26%
				Q:Depth, G: Depth	57.48% ± 0.08%	89.01% ± 0.13%	97.53% ± 0.41%	59.06% ± 0.20%

Table 5: One-stream network, BIWI: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

images from the *single-modal networks* in figure 27 from chapter 2.2, many differences are visible.



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 29: One-stream networks: Gradient images for BIWI with Resnet18 and softmax loss.

The RGB-based images (a) and (c) are less activated by certain surface regions, like the torso or head in comparison to the *single-modal networks* and are more activated by structural elements. It can be assumed, that the training procedure, to a certain level lowered the color dependence of the features. This is desirable for a common feature space as color features are not extractable in the depth modality. Nevertheless, the network is activated quite heavily by the colors of the upper leg region of the person and, hence, not completely focusing on the structure of the person. Also in the depth modality (images (b) and (d)) some differences to the single-modal network activations in figure 27 are apparent. Again the network is activated by the structure of the persons shape and a differentiation between torso and other parts of the body like arms are made. In comparison to figure 27 the activation by the torso shape is lower and the network is more focused on outer bounds of the person. Still the network is activated by different parts of the image for the depth and the RGB modality.

A visualization for the cross-modal query-gallery results with RGB as query can be found in the appendix in figure 45. The most important message of this image is, that the task the networks are solving in the cross-modal space is very challenging. Looking at figure 45, it gets clear, that it is also difficult for a human to classify the images correctly.

5.2.1.2 RobotPKU dataset

In table 6 the results of the one-stream network on the test set for RobotPKU are shown. For Resnet18 an average mAP of 9.34% for RGB as query and depth as gallery and 12.73% for depth as query and RGB as gallery are achieved. Similarly to the BIWI

dataset, the performance increases when changing to a Resnet50 architecture. Here, with RGB as query and depth as gallery an average mAP of 11.42% and with depth as query and RGB as gallery 14.19% are obtained.
In the single-modal tasks with Resnet18 an average mAP of 77.03% for RGB and 33.82% for depth are achieved. For the *single-modal network* with the same architecture these values were at *86.86%* and *38.65%*. Hence, a slight deterioration of this performance of around 9% and 5% was sensible. For the single-modal task in Resnet50 average mAPs of 79.00% and 38.34% are obtained. Again these values are slightly deteriorated by 8% and 5% in comparison to *87.11%* and *44.50%* in the *single-modal network*.

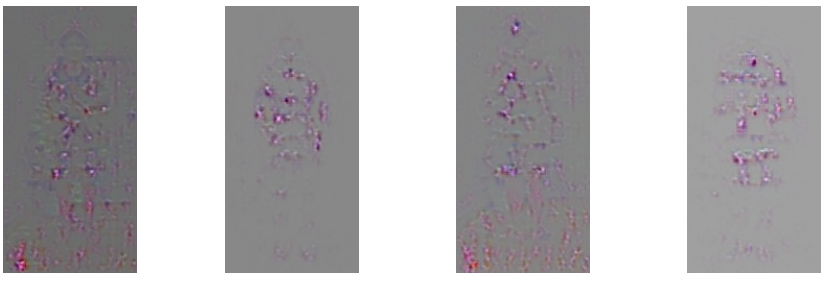
In figure 30 the activation maps of the one-stream networks for the RobotPKU dataset

Architect.	Datas.	Loss	Feature Extractor	Inference modality	R1	R5	R10	mAP
One-stream network	RobotPKU	Softmax	Resnet18	Q:RGB, G: Depth	10.06% ± 0.89%	34.08% ± 2.62%	52.25% ± 3.57%	9.34% ± 0.67%
				Q:Depth, G: RGB	11.17% ± 1.59%	36.96% ± 3.87%	54.87% ± 4.15%	12.73% ± 1.48%
				Q: RGB, G: RGB	75.52% ± 0.85%	94.26% ± 0.44%	97.40% ± 0.22%	77.03% ± 0.88%
				Q: Depth, G: Depth	34.68% ± 1.47%	64.87% ± 3.07%	78.65% ± 2.98%	33.82% ± 1.55%
			Resnet50	Q:RGB, G: Depth	11.92% ± 0.63%	38.13% ± 1.01%	57.34% ± 2.14%	11.42% ± 0.52%
				Q:Depth, G: RGB	12.48% ± 1.01%	38.51% ± 1.51%	56.77% ± 0.85%	14.19% ± 1.37%
				Q: RGB, G: RGB	77.27% ± 4.11%	94.92% ± 1.75%	97.62% ± 0.71%	79.00% ± 4.03%
				Q: Depth, G: Depth	38.52% ± 4.95%	69.23% ± 6.31%	82.19% ± 4.68%	38.34% ± 5.38%

Table 6: One-stream network, RobotPKU: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

are presented. The corresponding images from the original modalities were shown in figure 21.
Similar to the BIWI dataset the activation in the RGB modality is highly different to the one found in the *single-modal network* in figure 34. For the RGB modality, less influence of the salient regions, like head and feet are visible as it was the case in the *single-modal network*. The activation is much more cluttered and less easy to interpret. On the other hand, the activations in the depth modality (image (b) and (d)) are not much differing from the ones found in the single-modal network. Therefore, it can be concluded, that the features extracted from the RGB modality are capturing more of the general shape of the person than before. The fact, that almost no performance loss was found in the single-modal task for RGB, shows that an extraction of more shape-based features is not destructive for performance in the single-modal network.

Also for the one-stream network for the RobotPKU dataset some visualization can



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 30: One-stream networks: Gradient images for RobotPKU with Resnet18 and softmax loss.

be found in the appendix. In figure 49 an exemplary query-gallery result is visualized. Again, the difficulty of the task gets visible in the image.

5.2.1.3 SYSU RGB-IR dataset

The results for the SYSU dataset with one-stream network on the test set are reported in table 7. For the cross-modal task with a one-stream network trained with Resnet18 with RGB as query and infrared as gallery a mAP of 9.75% and with infrared as query and RGB as gallery a mAP of 12.64% is obtained. Again the one-stream network profits from a deeper network architecture and with Resnet50 a mAP for RGB as query and infrared as gallery of 14.19% and for depth as query and infrared as gallery of 18.98% is obtained.

Architect.	**Datas.**	**Loss**	**Feature Extractor**	**Inference modality**	**R1**	**R5**	**R10**	**mAP**
One-stream network	**SYSU**	**Softmax**	**Resnet18**	Q:RGB, G: Infrared	9.64%	26.69%	37.71%	9.75%
				Q:Infrared, G: RGB	12.11%	31.99%	44.45%	12.64%
				Q: RGB, G: RGB	62.57%	83.16%	88.49%	63.10%
				Q: Infrared, G: Infrared	51.13%	77.84%	85.96%	52.47%
			Resnet50	Q:RGB, G: Infrared	13.57%	36.39%	50.00%	14.19%
				Q:Infrared, G: RGB	18.40%	43.76%	58.04%	18.98%
				Q: RGB, G: RGB	73.25%	90.56%	94.44%	74.84%
				Q: Infrared, G: Infrared	61.62%	86.56%	93.57%	63.33%

Table 7: One-stream network, SYSU: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

For the single-modal task Resnet18 obtains a mAP of 63.10% for RGB and 52.47% for infrared. For the same architecture the *single-modal network* obtained a mAP of *68.38%* for RGB and *63.24%* for infrared. Therefore, for RGB a slight deterioration of about 5% and a more significant deterioration for infrared of around 11% is found. For the single-modal task with the one-stream network in Resnet50 mAPs of 74.84% for RGB and 63.33% for infrared images are obtained. Again this corresponds to a slight deterioration to the performance in the *single-modal networks* with *76.09%* and *69.91%*.

For the SYSU dataset no visualization of the activating regions in the images are made. The reason for this is that the dataset is only used as a test for the generalization of the techniques. The main focus of this work is on re-identification between RGB and depth and, hence, the BIWI and the RobotPKU datasets. Exemplary visualization of query-gallery set results for the cross-modal task with RGB as query can be found in the appendix in figure 41.

5.2.1.4 Discussion

After the evaluation of the results for the one-stream neural networks on the relevant datasets, several findings can be summarized and the advantages and disadvantages of the methods can be discussed.

Firstly, it is apparent, that the method is very powerful in terms of sensing in the single-modal tasks. For all datasets the results were only slightly deteriorated in comparison to the the results from the networks which were optimized solely in a *single modality*. This shows, that an one-stream architecture is inherently capable of handling a mixed input of two different modalities.

Nevertheless, for one-stream networks the results in cross-modal sensing were highly inferior to the performance in the single-modal task. For the BIWI dataset the best mAP of 23.75% was reached in sensing with depth as query and RGB as gallery, which is 35% lower than for sensing in the more difficult single-modal task which is in this case the depth modality. For the RobotPKU dataset, the best mAP was achieved for sensing with depth as query and RGB as gallery with an mAP of 14.2%. This is around 20% lower than the performance in the more difficult single-modal task, which is again depth. In

SYSU the best mAP was reached for sensing with infrared as query and RGB as gallery with 18.98%. This is around 45% percent lower than sensing in the individual modalities with the same network.

Analyzing these results it can be concluded, that the one-stream network is not optimal for cross-modal sensing. Most probable, a reason for that is, that in the optimization no explicit constraint for the cross-modal tasks is existing. A potential reason for this could be the optimization with a softmax loss. As described in 2.2.2.2, there is no guarantee for a successful generalization for zero-shot learning as the loss does not include a direct optimization of the embeddings.

Nevertheless, the gradient images from re-identification between depth and RGB suggested, that the activations of depth and RGB images are much closer together than for networks trained separately in these modalities. Nevertheless, still noticeable differences between the modalities are apparent.

5.2.2 Zero-padding neural network

The zero-padding network was presented in section 3.2. The network architecture contains guidance on modality1-specific, modality2-specific and shared nodes and, hence, is considered a kind of two-stream neural network. The idea for the zero-padding network was introduced by Wu et al. [83].

The implementation in this work is slightly differing to the one presented by Wu et al. First, the architecture will be Resnet18 and Resnet50 instead of Resnet6 used by the authors to enable a comparability to the other methods presented in this work. Second, instead of having a two-channel input with one channel allocated to a specific modality, a three-channel input was retained. In this case, one channel was always zero-padded, while the other two channels were allocated to the two modalities. The reason for this procedure is, that for three input channels well initialized models exist and, therefore, less pitfalls in retraining the models are apparent. The general idea of the zero-padding networks with domain-specific and shared nodes is not limited by this implementation.

5.2.2.1 BIWI

The results of the zero-padding network architecture for the BIWI dataset are displayed in table 8.

The results for the cross-modal task are very low. For Resnet18 an average mAP of 5.65% and 6.52% for the two tasks are obtained. For Resnet50 the average mAP is at 5.02% and 7.60% for a switching gallery and query set. These values are significantly lower than the results which were obtained for the cross-modal task for the one-stream network (section 5.2.1).

Architect.	Datas.	Loss	Feature Extractor	Inference modality	R1	R5	R10	mAP
Zero-padding network	BIWI	Softmax	Resnet18	Q:RGB, G: Depth	4.77% ± 0.71%	20.75% ± 3.08%	39.26% ± 7.78%	5.65% ± 0.71%
				Q:Depth, G: RGB	3.21% ± 0.42%	16.53% ± 2.34%	32.57% ± 4.12%	6.52% ± 0.61%
				Q:RGB, G: RGB	76.75% ± 4.61%	95.89% ± 1.20%	99.07% ± 0.29%	79.18% ± 4.47%
				Q:Depth, G: Depth	26.42% ± 4.63%	54.31% ± 5.18%	71.33% ± 2.17%	25.96% ± 4.54%
			Resnet50	Q:RGB, G: Depth	3.89% ± 0.34%	16.21% ± 1.83%	31.60% ± 2.92%	5.02% ± 0.40%
				Q:Depth, G: RGB	5.00% ± 1.74%	19.88% ± 2.26%	34.01% ± 1.19%	7.60% ± 1.47%
				Q:RGB, G: RGB	81.46% ± 3.06%	97.31% ± 0.79%	99.00% ± 0.49%	83.73% ± 2.84%
				Q:Depth, G: Depth	30.14% ± 0.15%	58.93% ± 1.16%	76.36% ± 2.59%	30.30% ± 0.24%

Table 8: Zero-padding network, BIWI: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

For the single-modal tasks the average mAP for Resnet18 is at 79.18% for RGB and 25.96% for depth. Within the *single-modal networks* these values were at *94.46%* and *58.38%*. Therefore, a noticeable performance difference is sensible. For Resnet50 a similar effect is visible. For sensing in the single-modal task in RGB the average mAP is deteriorated from *95.68%* to 83.73% and for sensing in the single-modal task in depth it is deteriorated from *61.44%* to 30.30%. Especially for the depth modality the differences are quite significant, but also a visible deterioration for the RGB modality took place. Most probably, the performance reduction can be explained with the fact, that the modalities are compressed into one channel instead of being introduced as three channels. As a matter of fact, the performance in the single-modal tasks decreases significantly. This effect is further enhanced for the cross-modal tasks where a very low performance is achieved.

5.2.2.2 RobotPKU

The results of the zero-padding network for the RobotPKU dataset can be seen in table 9. Again very low average mAPs in comparison to the one-stream networks are obtained. For Resnet18 the average mAP for the two cross-modal tasks lies at 4.76% and 4.98%. For Resnet50 the performance is even further deteriorated with 3.64% and 4.84%. These values lie very close to random guessing. Hence, it can be assumed that the zero-padding network in this architecture is not applicable for cross-modal sensing in the RobotPKU dataset.

Architect.	Datas.	Loss	Feature Extractor	Inference modality	R1	R5	R10	mAP
Zero-padding network	RobotPKU	Softmax	Resnet18	Q:RGB, G: Depth	4.62% ± 0.66%	19.24% ± 2.47%	34.27% ± 3.25%	4.76% ± 0.47%
				Q:Depth, G: RGB	3.54% ± 0.72%	17.07% ± 1.78%	32.28% ± 1.93%	4.98% ± 0.59%
				Q:RGB, G: RGB	60.85% ± 0.39%	87.37% ± 0.97%	93.68% ± 0.82%	62.82% ± 0.32%
				Q:Depth, G: Depth	20.98% ± 1.31%	39.89% ± 2.17%	53.77% ± 2.19%	19.71% ± 1.40%
			Resnet50	Q:RGB, G: Depth	3.42% ± 1.01%	15.18% ± 3.28%	29.25% ± 4.57%	3.64% ± 0.77%
				Q:Depth, G: RGB	3.04% ± 2.05%	14.61% ± 4.41%	29.60% ± 5.65%	4.84% ± 1.61%
				Q:RGB, G: RGB	57.96% ± 5.83%	84.07% ± 4.27%	91.80% ± 2.14%	59.99% ± 6.24%
				Q:Depth, G: Depth	18.67% ± 3.78%	35.83% ± 5.68%	47.88% ± 6.31%	17.60% ± 3.85%

Table 9: Zero-padding network, RobotPKU: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

Also in the single-modal tasks again a deterioration took place. The average mAP for Resnet18 for sensing in pure RGB is at 62.82%, while in the *single-modal networks* an average mAP of *86.86%* was achieved. For pure depth the deterioration was from 38.65 to *19.71%*. For Resnet50 both performance measures within the zero-padding network were even lower than for Resnet18, even though in the *single-modal networks* a higher performance was achieved with Resnet50.

5.2.2.3 SYSU-IR

The results of the zero-padding network in the SYSU-IR dataset are provided in table 10. Again comparably low mAPs in the cross-modal tasks are achieved. In the single-modal tasks the measures are significantly inferior to the results in the *single-modal networks*. The deterioration for RGB is at around 30% for Resnet18 and Resnet50. For depth a loss of 10% and 30% can be observed for Resnet18 and Resnet50, respectively. For cross-modal sensing the best results are obtained for Resnet50 with a mAP of 9.63% for RGB images as query and infrared images as gallery and 8.66% for the reverse.

Architect.	Datas.	Loss	Feature Extractor	Inference modality	R1	R5	R10	mAP
Zero-padding network	SYSU	Softmax	Resnet18	Q:RGB, G: Depth	8.7%	28.2%	42.7%	8.8%
				Q:Depth, G: RGB	8.4%	25.2%	38.7%	8.5%
				Q:RGB, G: RGB	50.3%	79.1%	88.0%	28.3%
				Q:Depth, G: Depth	28.8%	51.1%	64.9%	28.3%
			Resnet50	Q:RGB, G: Depth	10.17%	30.02%	45.16%	9.63%
				Q:Depth, G: RGB	8.83%	26.23%	39.3%	8.66%
				Q:RGB, G: RGB	55.3%	83.32%	91.09&	48.75%
				Q:Depth, G: Depth	22.53%	47.64%	61.36%	12.93%

Table 10: Zero-padding network, SYSU: Performance in test set. All possibilities for populating Query (Q) and Gallery (G) are reported.

5.2.2.4 Discussion

It was shown, that the results for the zero-padding algorithm were generally inferior in all dimensions in comparison to the results with the one-stream network. This contradicts the findings of Wu et al. [83]. The reason for the bad performance can most likely be found in the architecture of the zero-padding network. Here, only one channel per modality is available. Therefore, the information in RGB is reduced and also for depth the network has less parameters to extract meaningful features. The most probable explanation for the different findings to Wu et al. [83] can be found in the different network architecture. Resnet18 and Resnet50 architecture are deeper models than the Resnet6 architecture used in [83]. Therefore, the differentiation in modality-specific and modality-shared nodes in zero-padding did not add learning capabilities to the network architecture.

5.2.3 Cross-modal distillation network

The cross-modal distillation network is considered the main contribution of this work. The architecture and training procedure for this network was presented in section 3.3. The cross-modal distillation network is trained in a two-step procedure. The first step is an optimization in single-modalities. The results for this step were shown in section 5.1. In step II of the method the knowledge obtained in step I is transferred to the corresponding second modality by hands of a cross-distillation step.

In the following for the cross-modal distillation method all possible combination for the knowledge transfer will be investigated. Therefore, at maximum eight models have to be investigated for each of datasets. On the highest level this is the transfer from RGB to depth and the transfer from depth to RGB. In each of these, Resnet18 and Resnet50 both trained with softmax loss and triplet loss in the baseline have to be analyzed.

Afterwards, a view on the influence of the choice of the embedding layer will be made for the most successful methods.

5.2.3.1 BIWI RGBD-ID

In table 11 the results for the cross-modal distillation networks with baseline networks trained with softmax loss are shown. As the networks are based on the single-modal networks presented in chapter 5.1 the performance in one of the two single-modal tasks is always identical to the results from the baseline network. Hence, in table 11 in (a) and (b) the depth modality has the same performance as in table 2 and for (c) and (d) the RGB modality has the same performance as in table 2 for the corresponding baseline

network. Taking a detailed look at the table several interesting findings can be named.

Datas.	Basel. Loss (St. I)	Distil. Loss (St. II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
BIWI	Softmax	MSE	Depth to RGB	(a) R18	Q:RGB, G: D	21.93% ± 1.93%	57.70% ± 2.59%	78.85% ± 1.87%	22.29% ± 2.15%
					Q:D, G: RGB	27.72% ± 2.14%	65.35% ± 1.52%	81.63% ± 1.28%	27.95% ± 2.91%
					Q:RGB, G:RGB	59.78% ± 2.55%	89.06% ± 1.32%	97.62% ± 0.67%	62.68% ± 2.57%
					Q:D, G:D	55.07% ± 0.51%	89.46% ± 0.90%	98.17% ± 0.28%	56.30% ± 0.59%
				(b) R50	**Q:RGB, G: D**	**26.70% ± 5.16%**	**66.73% ± 4.31%**	**85.40% ± 2.81%**	**27.13% ± 4.94%**
					Q:D, G: RGB	**29.78% ± 4.14%**	**70.43% ± 2.29%**	**89.05% ± 1.36%**	**30.94% ± 3.72%**
					Q:RGB, D:RGB	61.78% ± 2.11%	87.29% ± 0.56%	96.19% ± 0.14%	63.88% ± 2.02%
					Q:D, G:D	**59.54% ± 0.51%**	**90.41% ± 0.17%**	**97.88% ± 0.12%**	**60.99% ± 0.81%**
			RGB to Depth	(c) R18	Q:RGB, G: D	5.25% ± 1.28%	19.93% ± 0.20%	35.19% ± 1.97%	6.09% ± 0.91%
					Q:D, G: RGB	6.55% ± 0.81%	29.13% ± 2.80%	51.58% ± 4.66%	11.56% ± 0.55%
					Q:RGB, D:RGB	**94.88% ± 0.94%**	**99.80% ± 0.07%**	**99.98% ± 0.02%**	**95.87% ± 0.59%**
					Q:D, G:D	34.59% ± 3.54%	65.06% ± 5.79%	82.65% ± 5.34%	34.53% ± 3.89%
				(d) R50	Q:RGB, G: D	6.40% ± 0.90%	24.57% ± 3.41%	42.38% ± 4.90%	6.91% ± 0.72%
					Q:D, G: RGB	7.78% ± 1.34%	30.04% ± 4.53%	49.86% ± 4.76%	11.90% ± 1.50%
					Q: RGB, D: RGB	93.60% ± 0.77%	99.71% ± 0.19%	99.96% ± 0.03%	94.79% ± 0.77%
					Q:D, G:D	33.11% ± 2.45%	64.54% ± 3.93%	81.80% ± 3.34%	32.87% ± 2.53%

Table 11: BIWI: Results for cross-modal distillation networks, Baseline loss (Step I) is Softmax loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). Reported are all possibilities to populate Query (Q) and Gallery (G) with RGB and depth (D)

Data.	Basel. Loss (St. I)	Distil. Loss (St. II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
BIWI	Triplet	MSE	Depth to RGB	(c) R18	**Q:RGB, G: D**	**25.54% ± 2.10%**	**70.20% ± 3.48%**	**93.07% 1.81%**	**27.92% ± 1.74%**
					Q:D, G: RGB	**26.47% ± 3.12%**	**71.86% ± 2.14%**	**93.98% ± 1.31%**	**28.13% ± 3.07%**
					Q:RGB, G:RGB	57.27% ± 2.42%	87.72% ± 1.90%	97.21% ± 0.88%	59.21% ± 2.68%
					Q:D, G:D	**61.24% ± 2.55%**	**94.12% ± 0.77%**	**99.47% ± 0.17%**	**62.69% ± 2.72%**
				(d) R50	Q:RGB, G: D	22.35% ± 4.16%	63.17% ± 6.82%	87.78% ± 5.03%	24.52% ± 3.66%
					Q:D, G: RGB	23.71% ± 4.40%	65.72% ± 6.17%	89.99% ± 4.37%	25.53% ± 3.89%
					Q:RGB, G:RGB	46.27% ± 4.08%	79.23% ± 3.66%	93.01% ± 2.34%	47.71% ± 4.25%
					Q:D, G:D	54.97% ± 1.08%	91.35% ± 0.57%	99.10% ± 0.16%	55.99% ± 1.15%
			RGB to Depth	(e) R18	Q:RGB, G: D	7.97% ± 0.58%	33.08% ± 2.11%	58.86% ± 4.75%	8.94% ± 0.55%
					Q:D, G: RGB	7.56% ± 1.35%	33.93% ± 2.55%	57.14% ± 2.68%	13.07% ± 1.03%
					Q:RGB, G:RGB	**93.47% ± 1.67%**	**99.74% ± 0.21%**	**99.99% ± 0.01%**	**94.50% ± 1.35%**
					Q:D, G:D	26.84% ± 2.27%	64.10% ± 2.81%	84.09% ± 2.61%	25.36% ± 2.74%
				(f) R50	Q:RGB, G: D	6.63% ± 0.95%	29.74% ± 1.51%	54.66% ± 2.77%	7.95% ± 0.78%
					Q:D, G: RGB	7.43% ± 1.79%	30.96% ± 3.12%	55.28% ± 3.45%	12.86% ± 1.61%
					Q:RGB, G:RGB	92.12% ± 1.86%	99.61% ± 0.31%	99.93% ± 0.12%	93.50% ± 1.53%
					Q:D, G:D	25.92% ± 2.12%	63.03% ± 2.65%	83.63% ± 2.50%	24.51% ± 2.10%

Table 12: BIWI: Results for cross-modal distillation networks, Baseline loss (Step I) is Triplet loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). All possibilities to populate Query (Q) and Gallery (G) with RGB and depth (D) are reported.

First, the performance in the cross-modal task is significantly better, when the knowledge is transferred from a network optimized within the depth modality to the RGB modality ((a) and (b) in table 11) than for a network optimized within the RGB modality and transferred to the depth modality ((c) in table 11). The performance difference in average mAP is bigger than 18% for Resnet50, which is very significant. For the transfer the usage of Resnet50 pushed the performance in comparison to the usage of Resnet18.

A second very interesting finding is that the performance in the modality the knowledge was transferred to can be better than the performance in the originally trained modality. For example, in (a) and (b) in table 11 the average mAP in RGB is at 62.7% and 63.88% for Resnet18 and Resnet50, while the performance in the starting networks for depth were at 56% and 61%. This effect was not observed in (c) and (d) where the transfer took place from RGB to depth. This finding indicates that the cross-distillation step generalizes well for the transfer from depth to RGB.

In table 12 the same evaluations are shown for the cross-modal distillation networks where the baseline was trained with triplet loss. In this table, similar observations can be made. Again, the transfer from depth to RGB is much more successful than the transfer from RGB to depth by bigger than 15% for the best performing model. For the transfer of baseline networks trained with triplet loss a deeper network (Resnet50) did not bring better results than the shallower Resnet18 architectures.

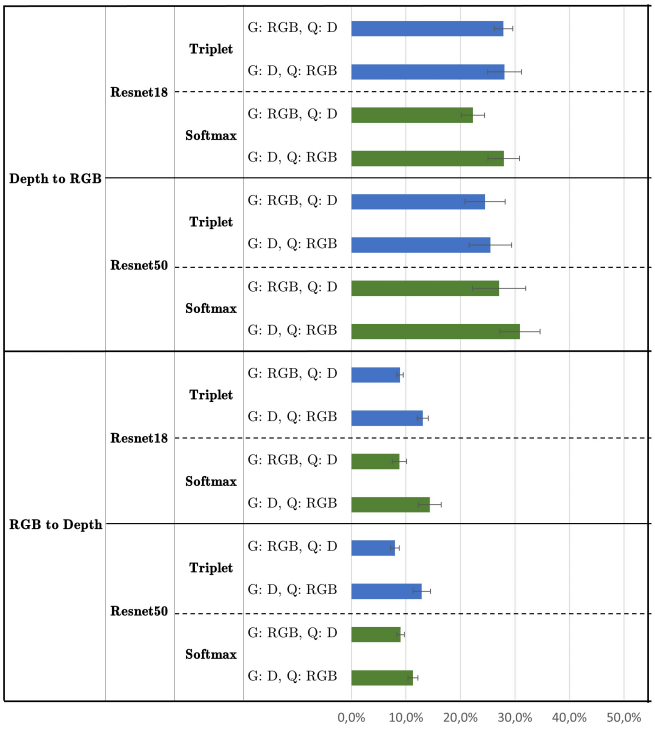
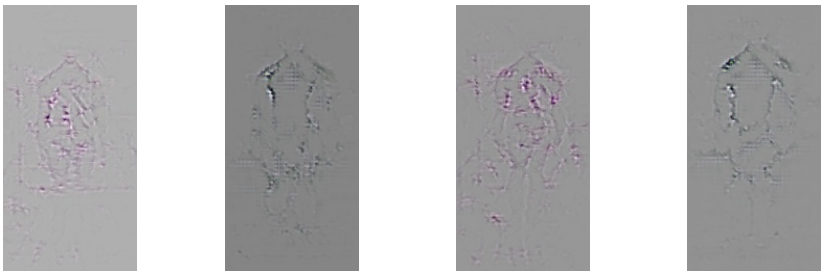


Figure 31: Overview over mAP performance for BIWI dataset with cross-distillation network. Only cross-modal tasks are reported.

Comparing table 11 and 12 it gets visible, that baseline models trained with triplet loss and with softmax loss, were both best performing in the cross-modal task when the knowledge was transferred from depth to RGB. A better visualization of the average mAPs of all transfer combinations can be seen in figure 31. For Resnet18 architectures the better result was obtained with a network trained with triplet loss. For Resnet50 a baseline model trained with Resnet50 was more suitable and this model was the overall best model.

In figure 32 the gradient visualization maps for the cross-distillation network from depth to RGB is shown. Following the architecture for the cross-distillation network, image (b) and (d) for the depth modality are the same as in figure 27 for the single-modal network. Comparing images (a) and (b) and images (c) and (d) it gets visible, that the cross-modal distillation was very successful. The gradient images from depth and from RGB are almost not differentiable. For all images, the activations are mainly based on the torso region, accompanied by the structure of the arms and upper legs. The results of the gradient images accompanied by the very good performance in the cross-modal task suggest, that the cross-distillation network was successfully deployed for the cross-modal tasks in the BIWI dataset.

Further visualizations for the cross-distillation network can be found in the appendix.



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 32: Cross-distillation networks: Gradient images for BIWI for Resnet18 and softmax loss baseline.

In figure 46 exemplary query-gallery images are shown.

5.2.3.2 RobotPKU

In the previous chapter several interesting findings for the cross-distillation network on the BIWI RGBD-ID dataset were made. In this chapter the results for the same tests for the RobotPKU dataset will be shown. This gives insights into the generalization of the findings from the BIWI dataset.

In tables 13 and 14 the performance for the cross-distillation network with softmax and triplet loss are shown. Again the performance in the baseline models is according to the performance in the models trained in section 5.1. This means, that for models (a) and (b) the performance in the single-modal task in depth is the same as for the single-modal task, while for (c), (d), (e), and (f) the performance in RGB is the same as in the single-modal task.

Dataset	Basel. Loss (St. I)	Distil. Loss (St. II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
RobotPKU	Softmax	MSE	Depth to RGB	(a) R18	Q:RGB, G: D	10.05% ± 2.03%	33.52% ± 5.42%	51.35% ± 5.96%	9.63% ± 1.63%
					Q:D, G: RGB	14.61% ± 2.43%	44.17% ± 5.70%	62.98% ± 5.77%	14.25% ± 2.51%
					Q:RGB, D:RGB	39.93% ± 5.07%	65.09% ± 6.68%	77.03% ± 6.19%	40.04% ± 5.50%
					Q:D, G:D	39.11% ± 0.72%	70.49% ± 0.43%	83.24% ± 0.20%	38.91% ± 0.84%
				(b) R50	**Q:RGB, G: D**	**18.36% ± 1.96%**	**49.85% ± 3.63%**	**68.53% ± 3.38%**	**17.38% ± 1.95%**
					Q:D, G: RGB	**18.82% ± 0.37%**	**51.76% ± 1.45%**	**70.55% ± 1.05%**	**17.90% ± 0.41%**
					Q:RGB, G:RGB	43.59% ± 1.33%	72.29% ± 0.50%	84.26% ± 0.59%	43.76% ± 1.44%
					Q:D, G:D	45.31% ± 0.53%	76.19% ± 0.66%	87.68% ± 0.57%	45.27% ± 0.36%
			RGB to Depth	(c) R18	Q:RGB, G: D	5.25% ± 1.28%	19.93% ± 0.20%	35.19% ± 1.97%	6.09% ± 0.91%
					Q:D, G: RGB	6.55% ± 0.81%	29.13% ± 2.80%	51.58% ± 4.66%	11.56% ± 0.55%
					Q:RGB, G:RGB	83.67% ± 0.44%	97.93% ± 0.32%	99.15% ± 0.08%	86.03% ± 0.44%
					Q:D, G:RGB	23.68% ± 1.35%	44.73% ± 2.73%	58.78% ± 3.40%	22.39% ± 1.59%
				(d) R50	Q:RGB, G: D	4.22% ± 1.78%	19.05% ± 5.37%	33.93% ± 6.34%	6.69% ± 1.92%
					Q:D, G: RGB	4.79% ± 1.37%	18.59% ± 2.96%	33.50% ± 4.29%	4.88% ± 1.04%
					Q:RGB, D:RGB	**84.25% ± 0.18%**	**97.67% ± 0.43%**	**98.94% ± 0.15%**	**86.70% ± 0.29%**
					Q:D, G:D	17.65% ± 2.10%	35.92% ± 4.86%	49.58% ± 6.14%	16.49% ± 2.21%

Table 13: RobotPKU: Results for cross-modal distillation networks, Baseline loss (Step I) is softmax loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). All possibilities to populate Query (Q) and Gallery (G) with RGB and depth (D) are reported.

In table 13 it gets visible, that again a transfer from depth to RGB (models (a) and (b)) was significantly more successful, than the other way round. The best results from Depth to RGB was achieved with Resnet50 and led to mAPs of 17.4% and 17.9%. This is more than 10% superior to the best transfor from RGB to depth. Nevertheless, the performance of all models in the cross-modal task with a best mAP of 17.9% is much

Dataset	Basel. Loss (St. I)	Distil. Loss (St. II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
RobotPKU	Triplet	MSE	RGB to Depth	(e) Resnet18	Q:RGB, G: D	6.09% ± 1.91%	24.10% ± 5.83%	41.70% ± 7.26%	6.26% ± 1.56%
					Q:D, G: RGB	6.89% ± 2.32%	24.92% ± 5.63%	41.90% ± 6.43%	10.38% ± 2.23%
					Q:RGB, G:RGB	90.51% ± 0.97%	99.38% ± 0.19%	99.56% ± 0.13%	91.85% ± 1.00%
					Q:D, G:D	19.30% ± 1.96%	47.54% ± 3.74%	66.16% ± 3.89%	17.34% ± 1.93%
				(f) Resnet50	Q:RGB, G: D	5.71% ± 0.78%	22.75% ± 1.30%	39.04% ± 1.90%	9.36% ± 0.74%
					Q:D, G: RGB	5.35% ± 0.92%	21.52% ± 2.16%	38.30% ± 3.13%	5.73% ± 0.91%
					Q:RGB, G:RGB	88.92% ± 3.66%	99.16% ± 0.13%	99.46% ± 0.08%	90.59% ± 3.28%
					Q:D, G:D	16.94% ± 1.03%	42.75% ± 3.20%	61.32% ± 4.08%	14.89% ± 1.00%

Table 14: RobotPKU: Results for cross-modal distillation networks, Baseline loss (Step I) is Triplet loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). All possibilities to populate Query (Q) and Gallery (G) with RGB and depth (D) are reported.

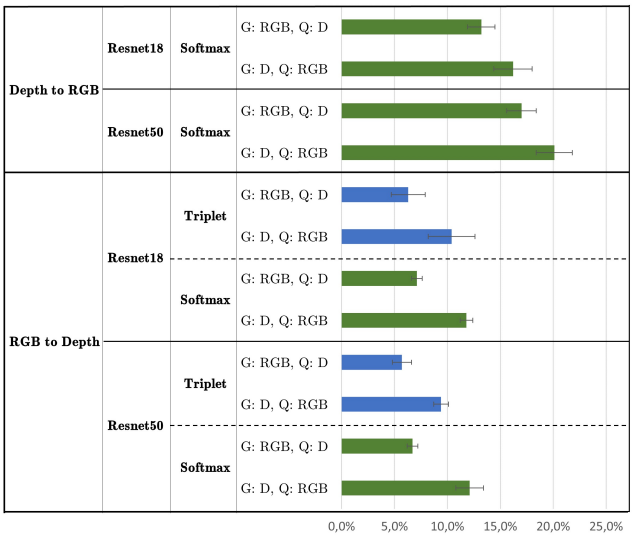


Figure 33: Overview over mAP performance for RobotPKU dataset with cross-distillation network.

lower in comparison to the performances which were found for the BIWI dataset. This shows the higher complexity of the task. For the transfers based on softmax the deeper Resnet50 network brought better results than the shallower Resnet18.

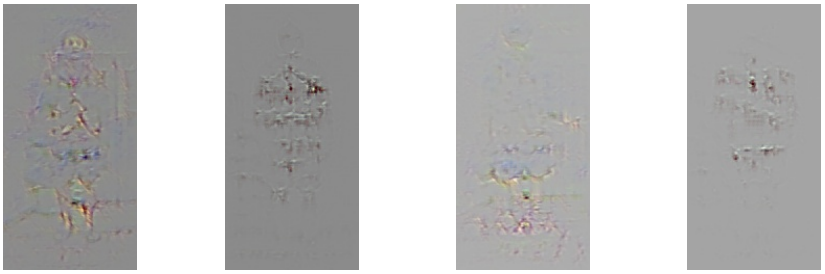
As described in chapter 5.1 it was not possible to successfully train models based on triplet loss for the depth modality. Therefore, in table 14 only networks for the transfer from RGB to depth are shown. It gets visible, that the average mAP reached by these models is significantly lower than the accuracies for the best models from softmax.

When comparing the cross-distillation steps based on triplet and softmax loss it gets visible, that the models based on softmax obtain the higher accuracies in the cross-modal tasks overall. Nevertheless, the triplet models are more successful to transfer the knowledge from RGB to depth than the corresponding models trained with softmax loss.

In figure 33 a visually faster to grasp overview over the average mAPs of the models is shown.

In figure 34 the obtained activation maps from softmax trained Resnet18 cross-modal distillation network are shown. Comparing the results to the gradients of the one-stream network (figure 30) it gets visible that very similar results are obtained. Again, the gradient images are closer to each other than for the single-modal networks (see figure

34), but they are not activated by exactly the same parts for depth and RGB. Therefore, the cross-modal distillation was successful, but not as impressing as for the BIWI dataset.



(a) Frontal image, RGB (b) Frontal image, Depth (c) Dorsal image, RGB (d) Dorsal image, RGB

Figure 34: Cross-distillation networks: Gradient images for RobotPKU with a baseline trained with Resnet18 and softmax loss.

Further visualizations for the RobotPKU cross-distillation network can be found in the appendix. In figure 50 a query-gallery result is shown.

5.2.3.3 SYSU-IR

For the SYSU IR dataset no coupled images are available. Therefore, to train the cross-modal distillation network instead of taking the embedding of the coupled image as groundtruth, the average over all images of one person instance was taken.

Dataset	Basel. Loss (Step I)	Distil. Loss (Step II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
SYSU	Softmax	MSE	Infrared to RGB	(a) R18	Q:RGB, G:I	7.52%	22.52%	34.53%	8.49%
					Q:I , G: RGB	5.48%	18.09%	28.54%	4.75%
					Q:RGB, G:RGB	13.78%	33.01%	44.73%	9.38%
					Q:I, G:I	61.76%	86.84%	93.11%	63.60%
				(b) R50	**Q:RGB, G:I**	**9.78%**	**27.80%**	**40.13%**	**10.87%**
					Q:I , G:RGB	**8.58%**	**26.00%**	**38.58%**	**7.67%**
					Q:RGB, G:RGB	20.25%	43.02%	55.25%	14.68%
					Q:I, G:I	**67.50%**	**90.73%**	**96.10%**	**69.44%**
			RGB to Infrared	(c) R18	Q:RGB, G:I	4.14%	12.89%	20.84%	3.96%
					Q:I , G: RGB	4.96%	17.37%	27.43%	5.76%
					Q:RGB, I:RGB	67.28%	87.40%	92.07%	68.38%
					Q:I, G:I	15.31%	29.46%	38.32%	14.05%
				(d) R50	Q:RGB, G:I	3.85%	12.72%	21.15%	3.46%
					Q:I , G: RGB	2.96%	12.23%	20.60%	4.06%
					G:RGB, Q:RGB	**75.06**	**91.35**	**94.76**	**76.09**
					Q:I, G:I	13.16%	26.04%	34.81%	12.19%

Table 15: SYSU: Results for cross-modal distillation networks, Baseline loss (Step I) is softmax loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). All possibilities to populate Query (Q) and Gallery (G) with RGB and Infrared (I) are reported

In table 15 and 16 the results for the cross-modal distillation method for the SYSU-IR datasets are shown.

Again transfers from RGB to infrared images and from infrared to RGB images were conducted. In chapter 5.1.3 it was shown, that sensing in the single modality of infrared was more difficult than sensing in the RGB modality for the dataset. In table 15 and 16 it gets visible, that the transfer from infrared to RGB worked better than the transfer

Dataset	Baseline Loss (St. I)	Distil. Loss (St. II)	Transfer Direction	FE	Inference modality	R1	R5	R10	mAP
SYSU	**Triplet**	**MSE**	**Infrared to RGB**	**(a) R18**	**Q:RGB, G:I**	**9.73%**	**30.08%**	**45.51%**	**9.74%**
					Q:I, G: RGB	**9.49%**	**31.43%**	**48.22%**	**11.51%**
					Q:RGB, G:RGB	13.99%	38.82%	55.54%	10.10%
					Q:I, G:I	**61.06%**	**88.34%**	**94.26%**	**62.66%**
			RGB to Infrared	**(c) R18**	Q:RGB, G:I	9.06%	27.81%	42.77%	9.31%
					Q:I, G: RGB	8.36%	28.68%	42.80%	10.34%
					Q:RGB, G:RGB	**74.00%**	**93.98%**	**97.52%**	**74.85%**
					Q:I, G:I	21.89%	47.13%	61.62%	21.38%

Table 16: SYSU: Results for cross-modal distillation networks, Baseline loss (Step I) is Triplet loss and distillation loss (Step II) is MSE. Variations in Transfer direction, Feature extractor (FE) between Resnet18 (R18) and Resnet50 (R50). All possibilities to populate Query (Q) and Gallery (G) with RGB and infrared (I) are reported.

from RGB to infrared. The best model for the cross-modal task were found within the baseline networks trained with the triplet loss.

Although the transfer from infrared to RGB worked better than the transfer from RGB to infrared, the overall transfer did not work well for both directions. This gets visible in the fact, that the single-modal performance in the transferred modality is very low for all cases. For example for the best performing model with softmax loss, which is Resnet50 trained from infrared to RGB has a mAP of 69.44% for the infrared modality and only a mAP of 14.68% for the RGB modality.

Several reasons for the bad performance of the cross-modal distillation network within the SYSU dataset can be found. First, no coupled images are available and, hence, one of the intrinsic ideas of the cross-modal distillation technique can not be applied. Second, the method with its distillation idea was designed for the transfer between depth and RGB. A finding is, that the asymmetrical relationship between depth and RGB is much different to the asymmetrical relationship between infrared images and RGB images.

A visualization of the query-gallery results of this network can be seen in the appendix in figure 42.

5.2.3.4 The embedding layer

Until now, in this work an embedding layer of size 128 was fixed for all analyses. This parameter was fixed to have a fair comparison between softmax and triplet loss. The size 128 was suggested in [45]. In fact, in most related literature the choice of this hyperparameter was not justified with experiments. To get better insights into a suitable embedding size we conducted experiments on the embedding size for the most successful models from the previous chapters. For BIWI these are the baseline models trained with Resnet50 and softmax loss as well as Resnet18 trained with triplet loss. For the RobotPKU a Resnet50 with softmax loss was re-evaluated. As explained in section 2.2 for softmax loss it is possible to take two different layers as the embedding. The penultimate layer (as described in equation (2)) can be varied in size and we chose to evaluate the embedding sizes 32, 128, 256, 512, 1024 and 2048. Additionally the classification layer before the softmax function (equation (3)) with size C of the classes will be evaluated.

In figure 35 the results for the BIWI dataset with a Resnet50 with softmax loss are shown with the varying embedding size. It gets visible, that the best cross-modal performance for the BIWI dataset is achieved with the classification layer embedding with size C. The next best model for the cross-modal tasks is the preliminary layer with

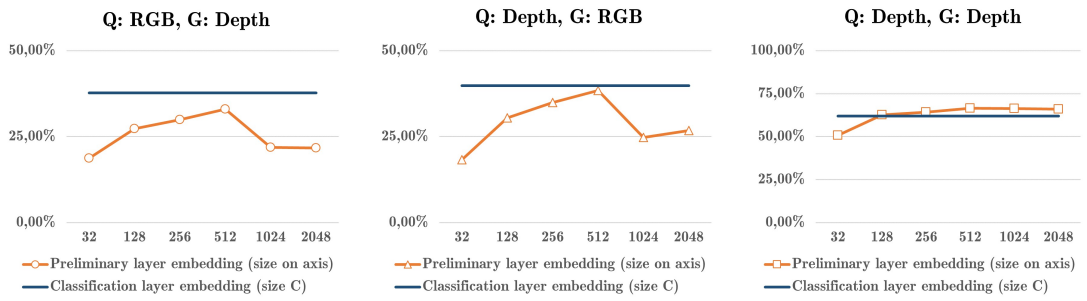


Figure 35: Analysis of influence of embedding layer and embedding size on the performance of the cross-modal distillation network with Resnet50 and *softmax loss* on the BIWI dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery (left), depth as query and RGB as gallery (middle) and single-modal performance in depth (right).

an embedding size of 512. It gets visible, that the performance of the penultimate layer is highly influenced by the size of the embedding. A clear maximum can be seen at a 512 feature embedding. Interestingly, for the single-modal task in depth the penultimate layer embeddings mostly outperformed the classification layer. Here, a bigger embedding size led to a slowly converging performance increase. Nevertheless, the classification layer was best suited for the cross-modal distillation task. The average mAP for the classification layer in the cross-modal tasks was 35.90%/38.31% for varying query and gallery population.

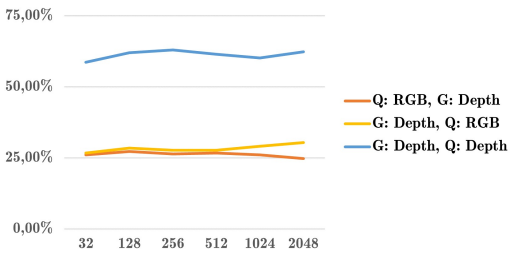


Figure 36: Analysis of influence of embedding size on the performance of the cross-modal distillation network with Resnet18 and *triplet loss* on the BIWI dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery, depth as query and RGB as gallery, and single-modal performance in depth in the same chart

The results for a varying embedding size for the triplet loss are shown in figure 36. Here only one definition of the embedding layer is existent and, hence, all statistics are shown in one figure. It gets visible, that the varying size of the embedding layer has very minor influence of the performance of the triplet loss based cross-modal distillation network. With a best cross-modal performance of 27.22%/30.42% the triplet loss based model is inferior to the softmax based model.

The results for a varying embedding size for RobotPKU can be seen in figure 37. Again the evaluation was made by hands of a Resnet50 with softmax loss. The classification layer embedding outperforms most of the embeddings from the penultimate layer within the cross-modal tasks. Nevertheless, in the RobotPKU dataset an embedding with size 256 for the penultimate layer gives better performance. Also in the single-modal task in depth a clear maximum for the embedding size can be seen at 256.

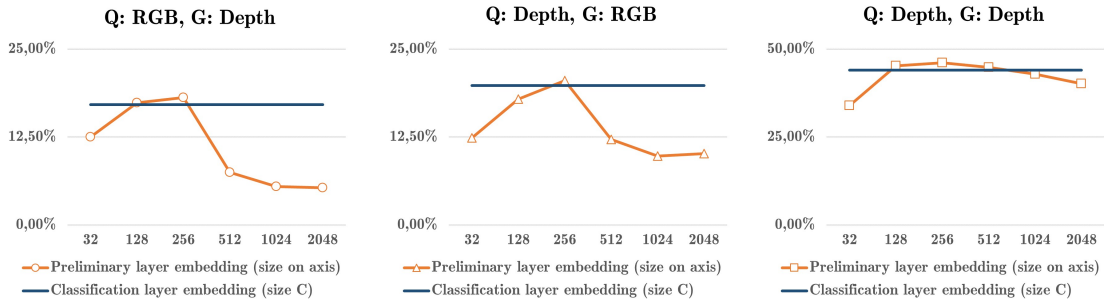


Figure 37: Analysis of influence of embedding layer and embedding size on the performance of the cross-modal distillation network with Resnet50 and softmax loss on the RobotPKU dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery (left), depth as query and RGB as gallery (middle) and single-modal performance in depth (right).

5.2.3.5 Discussion

After the inspection of the results in the individual datasets, the cross-distillation method can be discussed on a higher level. Therefore, it is necessary to revisit several aspects.

First, it is necessary to discuss the effectiveness of the cross-distillation direction. In the BIWI dataset, the best performance is reached for the transfer from the depth to the RGB modality. In this case, the performance difference to the other learning direction about 20 % (tables 11 and 12). For the RobotPKU dataset a similar behavior was observed. Again the transfer from depth to RGB was more effective than the transfer from RGB to depth. For the SYSU dataset the results for the correct direction for the knowledge transfer were not that clear. Still in softmax and triplet loss based cross-distillation models the transfer from infared to RGB was slightly more successful than the transfer from RGB to infrared. Considering all these results, a conclusion on the more effective transfer direction can be made. In all tested datasets the transfer from the 'weaker' modality, in terms of performance in the single-modal task (see tables 2, 3 and 7), to the stronger modality was more effective. In all cases, this means that the transfer from the RGB modality to the other modality was less effective. Two potential explanations for this behavior can be found. Firstly, the networks trained on RGB are very dependent on color features. Colors cannot be found in the corresponding other modalities (either infrared or depth). Therefore, it is not possible to transfer this knowledge to the corresponding modality. Secondly, most likely, the weaker modality contains information, which is also apparent in the RGB modality. For the infrared modality no colors are found but structural features and transitions are the same as in RGB. Therefore, it can up to a certain degree be considered a subspace of RGB. In depth, this is even more clear. The depth images are mainly capturing the structural appearance of a person. To a certain degree these features can be found in the RGB modality as well. This hypothesis is especially underlined by the findings in figure 32. It was shown, that it is possible to train a network in RGB which is very similarly activated as a network trained for the depth modality.

A second interesting point to look at is the performance in the single-modal task in the retrained modality. For BIWI in the cross-distillation with Resnet50 and softmax (table 11) the mAP performance in the retrained RGB modality is 63.88% and, therefore, even higher than the performance in the baseline depth network (60.99%). This shows, that a very meaningful transfer of knowledge took place. The knowledge transfer enabled,

that the sensing in the retrained modality can even be better than in the baseline. A similar result can be found in the RobotPKU dataset (see table 13). In this case for the baseline model with Resnet18 and softmax loss an average mAP of 40.04% in the RGB modality and 38.91% for the baseline in depth are achieved. In the SYSU dataset and several other cross-distillations, the effect was not observed. Nevertheless, the results are impressing as they clearly show that a knowledge transfer from depth to RGB is possible. On the other hand, the results also show, that the cross-distillation method does not seem to be universally applicable in cross-modal tasks as the results in the transfer from infrared to RGB was not successful and also a transfer from RGB to depth was not possible.

Additionally, we analyzed the influence of the embedding size on the performance of the cross-distillation network. It was shown, that the embedding size and embedding layer can have a significant influence on the performance of the cross-modal distillation network. For example for the BIWI dataset this difference was about 16% in average mAP between an embedding with the preliminary layer (2048D) and an embedding from the classification layer of the size of the training classes of 32 for a query from RGB and a gallery from depth. For BIWI dataset the best embedding was the classification layer with the size of the classes. For the RobotPKU dataset the best embedding was the penultimate layer with an embedding size of 256.

5.3 Comparison to state-of-the-art methods

After each of the three deep neural network methods was evaluated individually, in this chapter a overall comparison of the best performing version of the methods on the different datasets will be made.

To give an outright comparison, additionally to the neural network based models two conventional feature extractors, WHOS [35] and LOMO [21], were evaluated. Those feature extractors were used within a direct comparison in Euclidean space as well as after a metric learning step with XQDA [21].

Table 17: Average accuracy of state-of-the-art and proposed networks for different scenarios on the BIWI dataset. For results from [5] no detailed information on the evaluation procedure was given. As the single-gallery shot is used, this paper reports conservative accuracy indicators a comparison is still possible.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
R1 (%)	R5 ( )	R10 (%)	mAP (%)	R1 (%)	R5 (%)	R10 (%)	mAP (%)	
WHOS, Euclidean[35]	3.2	16.6	31.5	3.7	5.1	18.7	32.6	5.6
WHOS, XQDA[35]	8.4	31.7	50.2	7.9	11.6	34.1	51.4	12.1
LOMO, Euclidean[21]	2.8	16.4	32.5	4.8	3.3	15.6	29.8	5.6
LOMO, XQDA[21]	13.7	43.2	61.7	12.9	16.3	44.8	62.8	15.9
Eigen-depth HOG/SLTP, CCA [5]	8.4	26.3	41.6	-	6.6	27.6	45.0	-
Eigen-depth HOG/SLTP, LSSCDL [5]	9.5	27.1	46.1	-	7.4	29.5	50.3	-
Eigen-depth HOG/SLTP, Corr. Dict. [5]	12.1	28.4	44.5	-	11.3	30.3	48.2	-
Zero-padding network,[83] **Resnet50**	5.86 ± 2.18	25.85 ± 6.35	47.13 ± 8.06	7.28 ± 4.03	10.34 ± 2.68	38.91 ± 6.45	62.84 ± 11.48	9.77 ± 3.80
One-stream network,[83] **Resnet50**	15.68 ± 0.77	50.29 ± 1.18	75.65 ± 0.46	16.86 ± 0.87	19.82 ± 0.33	55.74 ± 0.83	78.92 ± 1.07	23.75 ± 0.30
Cross-modal distillation network, Resnet50 (ours), Emb. size 32 (C)	**34.87 ± 2.48**	**75.22 ± 2.42**	**93.93 ± 1.21**	**35.90 ± 2.37**	**36.29 ± 2.25**	**77.77 ± 2.21**	**94.44 ± 2.24**	**38.31 ± 2.18**

The results for the BIWI dataset can be seen in figure 17. For this dataset additionally to the three neural network techniques and the evaluated conventional approaches for this work, the findings of [5] were included. For the results of [5] no detailed information on the evaluation procedure was given. As the single-gallery shot is used for the rank accuracies, this paper reports conservative accuracy indicators and, hence, a comparison is still possible. A first finding from the table is that the zero-padding network is performing worse than the conventional methods. For example, the LOMO feature extractor combined with the XQDA feature learning obtains a mAP of 12.9%. The zero-padding network is significantly worse with an average mAP of 7.28%. The one-stream network is outperforming all conventional methods as well as the zero-padding network with an average mAP of 16.68%. This finding contradicts the finding of Wu et al. [83], where the zero-padding network was better as the one-stream network. The most probable explanation for the different results lies in the depth of the network. While Wu et al. [83] compared a one-stream network with the zero-padding network with a Resnet6, the used network for both evaluations in this work is Resnet50. Most probable, the higher capacity in Resnet50 led to the fact, that the zero-padding architecture is not adding value to the feature extraction anymore, while the one-stream network profits from it. Comparing the results from the one-stream network to the best performing cross-modal distillation network a significant difference is apparent. In average mAP the cross-modal distillation network outperforms the one-stream network by 19%/15% within the two cross-modal tasks. Therefore, the cross-modal distillation method can be considered the state-of-the-art model within the BIWI dataset for cross-modal person re-identification.

In table 18 the results for the RobotPKU dataset are shown. Again, similar obser-

Table 18: Average accuracy of state-of-the-art and proposed architecture for different scenarios on the RobotPKU dataset.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
	R1 (%)	R5 (%)	R10 (%)	mAP (%)	R1 (%)	R5 (%)	R10 (%)	mAP (%)
WHOS, Euclidean[35]	3.8	16.3	29.5	3.9	3.5	16.1	31.2	5.4
WHOS, XQDA[35]	10.0	31.8	49.8	8.2	9.8	31.0	48.0	9.8
LOMO, Euclidean[21]	3.6	15.0	28.0	3.9	3.7	15.3	28.7	4.9
LOMO, XQDA[21]	12.9	36.4	56.1	10.1	12.3	37.4	56.1	12.3
Zero-padding network,[83] **Resnet50**	7.76 ± 0.85	29.04 ± 2.57	47.79 ± 3.34	7.67 ± 0.59	6.57 ± 0.64	26.80 ± 2.14	45.62 ± 2.78	8.31 ± 0.56
One-stream network,[83] **Resnet50**	11.92 ± 0.63	38.13 ± 1.01	57.34 ± 2.14	11.42 ± 0.52	12.48 ± 1.01	38.51 ± 1.51	56.77 ± 0.85	14.19 ± 1.37
Cross-modal distillation network, Resnet50 (ours), Emb. size 256	**19.50 ± 0.99**	**50.11 ± 0.53**	**67.93 ± 0.69**	**18.13 ± 1.21**	**21.51 ± 1.12**	**54.90 ± 1.40**	**72.61 ± 0.95**	**20.52 ± 1.00**

vations can be made. The zero-padding network is inferior to conventional methods by several percentage points. The one-stream network again outperforms the zero-padding network as well as the conventional algorithms. In this case the difference between the best conventional model, LOMO with XQDA, and the one-stream network is in average mAP only 1.3%/1.8%. A significant performance boost is observed for the cross-modal distillation method. With an average mAP of 18.13%/20.52% it is by far the most efficient method for the task. It is 6.71%/6.33% superior to the one-stream network.

Table 19: State-of-the-art table for SYSU, including results from this work

Method	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
HOG+Euclidean	2.76	18.25	31.91	4.24	3.82	22.77	37.63	2.16	3.22	24.68	44.52	7.25	4.75	29.06	49.38	3.51
HOG+CRAFT	2.59	17.93	31.50	4.24	3.58	22.90	38.59	2.06	3.03	24.07	42.89	7.07	4.16	27.75	47.16	3.17
HOG+CCA	2.74	18.91	32.51	4.28	3.25	21.82	36.51	2.04	4.38	29.96	50.43	8.70	4.62	34.22	56.28	3.87
HOG+LFDA	2.33	18.58	33.38	4.35	3.82	20.48	35.84	2.20	2.44	24.13	45.50	6.87	3.42	25.27	45.11	3.19
LOMO+CCA	2.42	18.22	32.45	4.19	2.63	19.68	34.82	2.15	4.11	30.60	52.54	8.83	4.86	34.40	57.30	4.47
LOMO+CRAFT	2.34	18.70	32.93	4.22	3.03	21.70	37.05	2.13	3.89	27.55	48.16	8.37	2.45	20.20	38.15	2.69
LOMO+CDFE	3.64	23.18	37.28	4.53	4.70	28.23	43.05	2.28	5.75	34.35	54.90	10.19	7.36	40.38	60.33	5.64
LOMO+LFDA	2.98	21.11	35.36	4.81	3.86	24.01	40.54	2.61	4.81	32.16	52.50	9.56	6.27	36.29	58.11	5.15
Asymmetric FC layer network [83]	9.30	43.26	60.38	10.82	13.06	52.11	69.52	6.68	14.59	57.94	78.68	20.33	20.09	69.37	85.08	13.04
Two-stream network [83]	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
One-stream network [83]	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
One-stream network (Resnet18/ours)	16.13	56.27	72.17	18.92	19.76	60.38	75.32	13.73	19.65	64.45	79.69	30.30	25.02	70.12	83.24	21.70
Zero-padding network [83]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
BDTR reported [85]	17.01	55.43	71.96	19.66												
BDTR (ours, see appendix A) [85]	5.68	31.12	48.14	8.04	7.01	35.4	52.54	4.62	7.84	43.13	65.06	16.3	9.85	48.60	70.07	9.09
Cross-modal distillation network[85]	15.09	58.29	75.80	18.26	18.66	64.44	81.63	12.50	17.82	66.08	85.09	28.57	22.37	72.12	86.7	18.33
cmGAN [86]	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76

After the evaluation of the two datasets for the cross-modal person re-identification task in RGB and depth it can be concluded, that the cross-modal distillation method is the state-of-the-art for the task. Another aspiration of this work is to investigate how general the methods can be applied for cross-modal person re-identification. Therefore, also the SYSU dataset which consists of RGB and infrared images was investigated. The results can be seen in table 19.

When the SYSU dataset was presented in [83], the authors provided code for the evaluation of the datasets. Although the authors announced that a single-gallery shot setting will be used, their evaluation procedure is quite specific. The general definition of single-gallery shot settings is, that one image of each person in the gallery is used. Instead the authors decided to use one image of each person from each camera for the single-gallery shot evaluation. Therefore, the results are theoretically better than for a normal random single-gallery shot setting as it was used in this work. Hence, the earlier results from this work are brought into the evaluation setting of [83] to make the methods comparable. The resulting table can be seen in figure 19. Several interesting findings can be made. Firstly, the one-stream network trained with Resnet18 is superior

to the one-stream network trained by [83] with Resnet6. On top of that the Resnet18 one-stream network trained in this work is also superior to the zero-padding network which was reported by [83]. This result underlines the findings for the performance of the zero-padding network for the BIWI and RobotPKU dataset.

The results for the cross-modal distillation network are inferior to the results of the one-stream network for the SYSU dataset. Also the cmGAN [86] which was presented in mid-2018 is significantly outperforming all other reported methods. Hence, it can be concluded that the cross-modal distillation network is not generally applicable to all modality combinations for cross-modal person re-identification.

6 Conclusions

To sum up the findings of this work, in this conclusions section the research questions which were posed in the introduction (section 1) will be answered in detail. Additionally, the implications of the results for the intelligent vehicles domain will be discussed and some future directions for the research on cross-modal re-identification between depth and RGB will be given.

The first research question was *"How and how well can the cross-modal person re-identification task between RGB and depth be solved?"*

To answer this question two person datasets where Kinect depth images as well as RGB images are available were investigated. While the BIWI dataset [87] consists of cleanly captured and well aligned images, the RobotPKU dataset [88] is more noisy in terms of alignment of the images and image quality.

Several techniques were investigated to evaluate the feasibility of cross-modal person re-identification in these datasets. Hereby, the focus was on neural network techniques as this was suggested by the success of these methods in closely related work. In chapter 5.3 the focus on deep neural networks was justified as two of the three neural network techniques were superior to all evaluated conventional methods for both RGB-D datasets. The one-stream network which was suggested by Wu et al [83] outperformed all conventional methods and the zero-padding network by 3.96%/7.83% in average mAP with changing query and gallery for the BIWI dataset and by 1.32%/1.89% for the RobotPKU dataset. The cross-modal distillation network which was presented in this work was able to outperform the one-stream network in average mAP by additional 19.04%/14.56% for BIWI and 6.71%/6.33% for the RobotPKU dataset. Hence, the cross-modal distillation network is considered the state-of-the-art for cross-modal person re-identification between depth and RGB.

A reason for the superiority of the cross-modal distillation network in comparison to the other methods can be inferred from the activation maps of the different methods. In figure 38 the activation maps from the single-modal networks, one-stream network and cross-modal distillation network which were discussed in the corresponding chapters are summarized together. It gets visible that the single-modal networks in the different modalities are activated by very different parts of the images. In the RGB modality (images (a) and (c) in figure 38) the activations are mainly triggered by colors. In the depth modality (images (b) and (d)) the activations of the network can be found in the shape of the person. In the third column the corresponding activation images for the one-stream network are visualized. It gets visible that the activations are closely connected to the activations within the single-modal networks. Nevertheless there is a slight trend in the RGB images to be activated by the shape of the person. Finally, in the fourth column the activations of the cross-modal distillation network are shown. Here, clearly in both modalities similar features are extracted. To be more exact in both modalities features which were found in the single-modal networks in depth are activating the networks. This shows that the transfer of knowledge from depth to RGB was successful. As a consequence of the similar activation in both modalities an embedding to the same feature space is facilitated. Hence, the better accuracies in the cross-modal tasks can be explained.

Based on these results, the posed research question can be answered. The task of cross-modal re-identification can be solved with deep neural network structures. It was shown, that several approaches based on neural networks exist which can be successfully

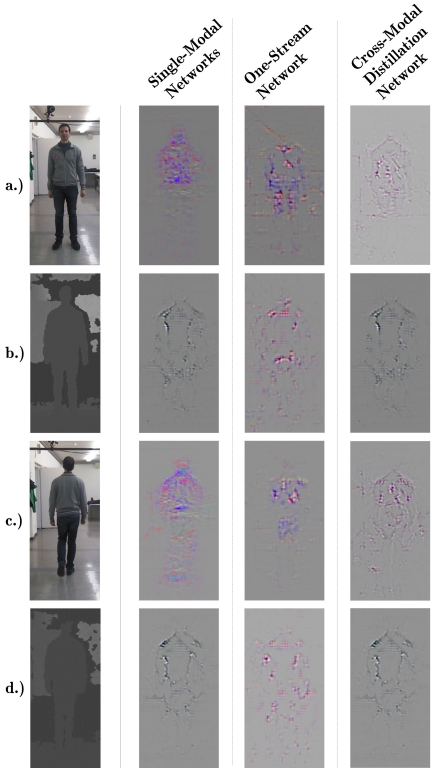


Figure 38: Comparison of activation maps for single-modal networks (2nd column), one-stream network (3rd column) and cross-modal distillation network (4th column) for the BIWI dataset. Original images in the left column.

deployed for the task. On top of that it was shown, that when exploiting the inherent relationship between depth and RGB even better results can be achieved.

The second research question of this work was *"Is it possible to develop a generic methodology for re-identification between modalities and how well does it generalize to different modality combinations, like near-infrared images and visible light images?"*

In this work additionally to the datasets with RGB and depth images a dataset with infrared and RGB images was investigated. For this dataset the cross-modal distillation method was not the most successful method tested in this work and also in related literature superior methods can be found. The one-stream network, which was significantly inferior to the cross-modal distillation method for the RGB-depth datasets, outperformed the cross-modal distillation method for this dataset. Therefore, it is difficult to find a conclusive answer on this research question. In general, the methods are designed such that they are theoretically generic for cross-modal re-identification tasks. Also the experimental results show that the neural network based methods are applicable on all kinds of dataset. Nevertheless, the performance of the different techniques in different datasets suggest, that it might be necessary to craft specific methods for the specific asymmetrical relationships of certain cross-modal re-identification task.

Overall the analysis in this paper showed that cross-modal person re-identification is a complex task, and the results in absolute numbers suggest that there is still room for improvement. In fact, the accuracies obtained in cross-modal re-identification (tables 17 and 18) are still significantly lower than the accuracies for single-modal re-identification

in the more difficult modality for re-identification (tables 2 and 3). This hints that further improvements will be possible. As this is one of the first works concerning the task it is possible to highlight some potential future directions and current problems in the domain.

First, it will be necessary to obtain bigger datasets to make research in re-identification between depth and RGB more attractive and give data-hungry methods based on deep neural networks the possibility to obtain higher accuracies. The publication of the SYSU-IR dataset in 2017 [83] pushed the interest in cross-modal person re-identification in RGB vs. infrared immensely [84, 85, 86]. A similar effect could be expected for cross-modal re-identification between RGB and depth. Therefore, the amount of persons contained in the datasets would have to rise from less than a hundred for the current datasets to at least the magnitude of several hundreds. Additionally, high-quality depth and RGB images will be necessary.

Second, for future research it will be important to expand the considered mode of depth. Especially for the needs in intelligent vehicles it will be necessary to evaluate the methods on sparse depth maps, as captured by lidars or radars. Therefore, completely new datasets with a high amount of tracked pedestrians and other street objects, will be needed.

Overall, this work approached the relevant problem of cross-modal re-identification in RGB and depth for surveillance applications as well as intelligent vehicles in a very effective way. The newly presented method brings the community closer to solve this difficult problem and the results help to understand the relation between RGB and depth better.

References

[1] http://psychologydictionary.org/, 15.04.2018

[2] S. Schumacher, de Perera, T. B., Thenert, J., von der Emde, G. “Cross-modal object recognition and dynamic weighting of sensory inputs in a fish,” Proceedings of the National Academy of Sciences. 2016.

[3] http://www.animaldiscoveryonline.com/elephant-nose-fish/, 13.04.2018

[4] Xiong, F. et al. "Person re-identification using kernel-based metric learning methods." European conference on computer vision. 2014.

[5] Zhuo, J., et al. "Person Re-identification on Heterogeneous Camera Network." CCF Chinese Conference on Computer Vision. 2017.

[6] Lohani, B., Chacko, S., Ghosh, S., Sasidharan, S. "Surveillance system based on Flash LiDAR." Proceedings of XXXII INCA International Congress on Cartography for Sustainable Earth Resource Management. Vol. 32. 2013.

[7] Sudhakar, P., Anitha Sheela, K., Satyanarayana, M. "Imaging Lidar system for night vision and surveillance applications." Advanced Computing and Communication Systems (ICACCS). 2017.

[8] https://www.springerprofessional.de/automatisiertes-fahren/sensorik/hyundai-stellt-autonomes-serienkonzept-vor/11080526, 25.09.2018

[9] Gong, S., et al. "The re-identification challenge." Person re-identification. 2014

[10] Lienhart, R., Maydt, J.. "An extended set of haar-like features for rapid object detection." International Conference on Image Processing. Vol. 1. IEEE, 2002.

[11] Gavrila, D. M. "Pedestrian detection from a moving vehicle." European conference on computer vision, 2000.

[12] Ren, S., et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[13] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems. 2016.

[14] Gong, S., et al., "Person re-identification". 2014.

[15] Li, W., Zhao, R., Wang, X. "Human reidentification with transferred metric learning." Asian Conference on Computer Vision. 2012.

[16] Li, W., Wang, X.. "Locally aligned feature transforms across views." in Conference on Computer Vision and Pattern Recognition. 2013.

[17] Li, W., Zhao, R., Xiao, T., Wang, X. “DeepReID: Deep filter pairing neural network for person re-identification,” in Conference on Computer Vision and Pattern Recognition. 2014.

[18] Zhao, R., Ouyang, W., Wang, X. “Learning mid-level filters for person re-identification,” in Conference on Computer Vision and Pattern Recognition. 2014.

[19] Goodfellow, I., et al. "Deep learning". Cambridge: MIT press, 2016.

[20] Zheng, L., Yang, Y., Hauptmann, A.. "Person re-identification: Past, present and future." arXiv 2016.

[21] Liao, S. and Hu, Y. and Zhu, X. and Li, S. Z. "Person re-identification by local maximal occurrence representation and metric learning". Computer Vision and Pattern Recognition. 2015.

[22] Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S. Z. "Salient color names for person re-identification," European Conference on Computer Vision. 2014.

[23] Gheissari, N., Sebastian, T. B., Hartley, R. "Person re-identification using spatio=temporal appearance". CVPR. 2006.

[24] Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M. "Person re-identification by symmetry-driven accumulation of local features." CVPR. 2010.

[25] Cheng, D.S., Cristani, M. , Stoppa, M., Bazzani, L., Murino, V. "Custom pictorial structures for re-identification". BMVC. 2011.

[26] Bhuiyan, A. and Perina, A. and Murino, V. "Person re-identification by discriminatively selecting parts and features". ECCVWK. 2014.

[27] Bak, S., Corvee, E., Bremond, F., Thonnat, M. "Boosted human re-identification using riemannian manifolds." Image and Vision Computing. 2012.

[28] Farenzena, M. and Bazzani, L., Perina, A., Murino, V., Cristani, M. "Person re-identification by symmetry-driven accumulation of local features" CVPR. 2010.

[29] Gray, D., Tao, H. "Viewpoint invariant pedestrian recognition with an ensemble of localized features." ECCV. 2008.

[30] Prosser, B., Zheng, W.-S., Gong, S., Xiang, T., Mary, Q. "Person re-identification by support vector ranking." BMVC, 2010.

[31] Schwartz, R., Davis, L. S. "Learning discriminative appearance-based models using partial least squares." Brazilian Symposium on Computer Graphics and Image Processing. 2009.

[32] Corvee, E., Bremond, F., Thonnat, M. et al. "Person re-identification using spatial covariance regions of human body parts." AVSS, 2010.

[33] Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B. "Person re-identification in multicamera system by signature based on interest point descriptors collected on short video sequences." Distributed Smart Cameras. 2008.

[34] Zheng, L., et al. "Scalable person re-identification: A benchmark." International Conference on Computer Vision. 2015.

[35] Lisanti, G., Masi, I., Bagdanov, A. D., Del Bimbo, A.."Person re-identification by iterative re-weighted sparse ranking." TPAMI. 2015.

[36] Liu, C. and Gong, S. and Loy, C. C. "On-the-fly feature importance mining for person re-identification." Pattern Recognition. 2014.

[37] Sanping Z., Jinjun W., Jiayun W., Yihong G., Nanning Z., “Point to Set Similarity Based Deep Feature Learning for Person Re-Identification,” CVPR. 2017.

[38] Figueira, D., Bazzani, L., Minh, H. Q., Cristani, M., Bernardino, A., Murino, V. "Semisupervised multi-feature learning for person re-identification." AVSS. 2013.

[39] Koestinger, M., Martin H., Paul W., Peter M. R., and Horst B. "Large scale metric learning from equivalence constraints." CVPR. 2012.

[40] Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. "Deep Metric Learning for Person Re-identification." CVPR. 2014.

[41] Ahmed, E., Jones, M., Marks, T.K. "An improved deep learning architecture for person re-identification." CVPR. 2015.

[42] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N. "Person re-identification by multi-channel parts-based CNN with improved triplet loss function." CVPR. 2016.

[43] Varior, R.R., Haloi, M., Wang, G. "Gated siamese convolutional neural network architecture for human re-identification." ECCV, 2016.

[44] Xiao, T., Li, H., Ouyang, W., Wang, X. "Learning deep feature representations with domain guided dropout for person re-identification." arXiv. 2016.

[45] Hermans, A., Beyer, L. and Leibe, B.. "In defense of the triplet loss for person re-identification." arXiv. 2017.

[46] Giuseppe L. , Iacopo M. , Alberto D. B., Matching People across Camera Views using Kernel Canonical Correlation Analysis”, Eighth ACM/IEEE International Conference on Distributed Smart Cameras, 2014.

[47] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.

[48] Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., Li, S. Z. "Embedding deep metric for person re-identification: A study against large variations." ECCV, 2016.

[49] Chen, W., Xiaotang, C., Jianguo, Z., and Huang, K. "Beyond triplet loss: a deep quadruplet network for person re-identification." CVPR. 2017.

[50] Ristani, E., Carlo, T. "Features for Multi-Target Multi-Camera Tracking and Re-Identification." arXiv 2018.

[51] Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L. "Joint learning of single-image and cross-image representations for person re-identification." Computer Vision and Pattern Recognition. 2016.

[52] Wang, F., et al. "Joint learning of single-image and cross-image representations for person re-identification." Computer Vision and Pattern Recognition. 2016.

[53] Varior, R.R., Haloi, M., Wang, G.. "Gated siamese convolutional neural network architecture for human re-identification." European Conference on Computer Vision. 2016.

[54] Li, D., Chen, X., Zhang, Z., Huang, K. "Learning deep context-aware features over body and latent parts for person re-identification." CVPR. 2017.

[55] Zheng, Z., Liang Z., and Yi Y. "A discriminatively learned cnn embedding for person reidentification." Transactions on Multimedia Computing, Communications, and Applications. 2017.

[56] He, K., et al. "Deep residual learning for image recognition." Computer vision and pattern recognition. 2016.

[57] Geng, M., Wang, Y., Xiang, T., Tian, Y. "Deep transfer learning for person re-identification." arXiv. 2016.

[58] Li, Y. J., Yang, F. E., Liu, Y. C., Yeh, Y. Y., Du, X., Wang, Y. C. F. "Adaptation and Re-Identification Network: An Unsupervised Deep Transfer Learning Approach to Person Re-Identification." arXiv. 2018.

[59] Wu, A., Zheng, W., Lai, J.. "Depth-based person re-identification." Asian Conference on Pattern Recognition (ACPR). 2015.

[60] Weinberger, Kilian Q., and Lawrence K. Saul. "Distance metric learning for large margin nearest neighbor classification." Journal of Machine Learning Research, 2009: 207-244.

[61] Barbosa, I.B., et al. "Looking beyond appearances: Synthetic training data for deep cnns in re-identification." Computer Vision and Image Understanding. 2017.

[62] Albiol, A., Oliver, J., Mossi, J. "Who is who at different cameras: people re-identification using depth cameras." Computer Vision, 2012.

[63] Munsell, B. C., Temlyakov, A., Qu, Wang, S. "Person identification using full-body motion and anthropometric biometrics from kinect videos." ECCV, 2012.

[64] Ioannidis, D., Tzovaras, D., Damousis, I. G., Argyropoulos, S., Moustakas, K. "Gait recognition using compact feature extraction transforms and depth information." Transactions on Information Forensics and security. 2007.

[65] Andersson, V., Dutra, R., Araujo, R. "Anthropometric and human gait identification using skeleton data from kinect sensor." Symposium on Applied Computing. 2014.

[66] Karianakis, N., et al. "Person Depth ReID: Robust Person Re-identification with Commodity Depth Sensors." arXiv. 2017.

[67] Haque, A., Alahi, A., Fei-Fei, L. “Recurrent attention models for depth-based person identification,” Conference on Computer Vision and Pattern Recognition. 2016.

[68] Schroff, F., Kalenichenko, D., Philbin, J. "Facenet: A unified embedding for face recognition and clustering." Conference on computer vision and pattern recognition. 2015.

[69] Barbosa, I.B., et al. "Re-identification with rgb-d sensors." European Conference on Computer Vision. 2012.

[70] Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G. "The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits." Journal of Visual Communication and Image Representation. 2014.

[71] Jain, A. K., and Stan, Z. Li. "Handbook of face recognition." New York: Springer. 2011.

[72] Wen, Y., et al. "A discriminative feature learning approach for deep face recognition." European Conference on Computer Vision. 2016.

[73] Sun, Y., Wang, X., Tang, X. "Deep learning face representation from predicting 10,000 classes." Conference on Computer Vision and Pattern Recognition. 2014.

[74] Taigman, Y., et al. "Deepface: Closing the gap to human-level performance in face verification." Conference on computer vision and pattern recognition. 2014.

[75] Chopra, S., Hadsell, R., LeCun, Y. "Learning a similarity metric discriminatively, with application to face verification." Conference on Computer Vision and Pattern Recognition. 2005.

[76] Wang, M., Deng, W. "Deep Visual Domain Adaptation: A Survey." Neurocomputing. 2018.

[77] Gupta, S., Hoffman, J., Malik, J. "Cross modal distillation for supervision transfer." Conference on Computer Vision and Pattern Recognition. 2016.

[78] Wu, A., Zheng, W., Lai, J. "Robust depth-based person re-identification." Transactions on Image Processing. 2017.

[79] Dalal, N., Triggs, B.. "Histograms of oriented gradients for human detection." Conference on Computer Vision and Pattern Recognition. 2005.

[80] Ojala, T., Pietikainen, M., Maenpaa, T. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." Transactions on pattern analysis and machine intelligence. 2002.

[81] Liao, S., et al. "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes." Conference on Computer Vision and Pattern Recognition. 2010.

[82] Springenberg, J. T., et al. "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806 (2014).

[83] Wu, A., et al. "RGB-infrared cross-modality person re-identification." 2017.

[84] Ye, M., et al. "Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification." 2018.

[85] Ye, M., et al. "Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking." IJCAI. 2018.

[86] Dai, P., et al. "Cross-Modality Person Re-Identification with Generative Adversarial Training." IJCAI. 2018.

[87] Munaro, M., et al. "One-shot person re-identification with a consumer depth camera." Person Re-Identification. 2014.

[88] Liu, H., Hu, L., Ma, L. "Online RGB-D person re-identification based on metric model update." CAAI Transactions on Intelligence Technology. 2017.

[89] Nesterov, Y.. "A method for unconstrained convex minimization problem with the rate of convergence O." Doklady AN USSR. Vol. 269. 1983.

[90] Kingma, D. P., Ba, J. "Adam: A method for stochastic optimization." arXiv. 2014.

Appendices

A BDTR

The BDTR two-stream neural network was discussed in chapter 2.3.3 and visualized in figure 15. The general idea is to have two input streams for a neural network which are specific to each modality. Subsequently, these individual streams are merged on a higher feature level. As the authors provided the codebase for their paper, the idea is not re-implemented and the implementation of the authors is used. Therefore, Alexnet is the feature extractor [85]. The main contribution of Ye et al. within the paper is the combination of the identity loss and the ranking loss as described in chapter 2.3.3. In the paper the authors do not define a validation set and the model is trained until convergence. This configuration was used for the evaluation in this work.

The authors published the algorithm based on the results on the SYSU-IR dataset. Hence, for this work the network was retrained with the parameters provided by Ye et al. [85] and evaluated in the same scheme.

The results reported in the paper and the results obtained by the retraining can be found in table 20. It is clearly visible, that a huge difference in the performance of the network is existent. The mAP differs around 11% whilst the values are 19.66% and 8.04%, respectively. Hence, most probably the authors evaluated the algorithm differently as it is shown in the published code. Even though not enough details are given in the paper, it is possible to identify two potential flaws. Firstly, there is the possibility that the authors used some kind of early-stopping criteria for the neural network training based on the test set. This leads to an overfitting on the test data, but at the same time pushes the performance of the algorithm. Secondly, the authors only used parts of the test set for evaluating the performance on the test set. The second hypothesis is reinforced by the provided code-base, where exactly 2060 images, which are not randomly shuffled are used for the evaluation in the test set.

Cross-modal	**R 1**	**R 10**	**R 20**	**mAP**
Reported in [85]	17.01	55.43	71.96	19.66
Reproduced result	5.68	31.12	48.14	8.04

Table 20: Results BDTR in SYSU

Due to the high uncertainty in terms of the correct training procedure of the two-stream network, at this point no further evaluation of the the architecture for the other datasets are made.

B Splits of the datasets

B.1 BIWI RGBD-ID

Design set (Train + Validation set):
0, 1, 4, 5, 6, 7, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20, 25, 26, 34, 35, 38, 39, 40, 43, 50, 56, 57, 58, 59, 61, 62, 65, 66, 67, 69, 70, 73, 74, 76, 77.
Test set:

2, 3, 8, 10, 14, 21, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 36, 37, 41, 42, 44, 45, 46, 47, 48, 49, 51, 52, 53, 54, 55, 60, 63, 64, 68, 71, 72, 75.

B.2 RobotPKU

Design set (Train + Validation set):
0, 2, 3, 15, 16, 18, 19, 20, 21, 22, 23, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 41, 43, 44, 45, 46, 47, 52, 54, 55, 58, 59, 60, 63, 66, 67, 68, 72, 73, 74, 77, 78, 80, 82, 83, 84, 87, 88.
Test set:
1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 24, 26, 38, 39, 40, 42, 48, 49, 50, 51, 53, 56, 57, 61, 62, 64, 65, 69, 70, 71, 75, 76, 79, 81, 85, 86, 89.

B.3 SYSU RGB-IR

Split predefined by the authors.
Train set:
0, 1, 3, 4, 6, 7, 10, 11, 12, 13, 14, 15, 17, 18, 19, 21, 28, 29, 34, 51, 52, 54, 55, 57, 58, 59, 60, 61, 63, 64, 69, 70, 71, 72, 73, 75, 76, 77, 78, 90, 91, 94, 97, 98, 106, 108, 109, 110, 112, 113, 114, 117, 118, 119, 120, 122, 123, 125, 126, 127, 130, 131, 132, 134, 135, 136, 139, 141, 142, 146, 148, 150, 153, 154, 155, 156, 157, 158, 159, 160, 162, 163, 164, 167, 168, 170, 173, 174, 176, 177, 178, 179, 180, 181, 182, 183, 185, 187, 188, 192, 193, 195, 196, 197, 198, 199, 200, 202, 204, 205, 207, 208, 210, 211, 212, 213, 215, 216, 217, 218, 219, 220, 221, 223, 224, 225, 226, 227, 229, 230, 233, 234, 239, 242, 243, 244, 245, 246, 247, 248, 249, 250, 253, 254, 255, 257, 259, 260, 261, 263, 264, 266, 267, 269, 270, 275, 276, 277, 278, 279, 280, 282, 283, 285, 286, 287, 288, 289, 291, 292, 293, 294, 295, 296, 297, 298, 303, 304, 305, 307, 308, 309, 310, 312, 313, 315, 316, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 331, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 529, 530, 531, 532
Validation set:
333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432
Test set:
5, 9, 16, 20, 23, 24, 26, 27, 30, 33, 35, 36, 39, 40, 41, 42, 43, 44, 48, 49, 50, 53, 62, 68, 74, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 92, 101, 103, 104, 105, 107, 111, 115, 116, 121, 124, 128, 129, 133, 137, 138, 149, 151, 161, 165, 166, 169, 171, 175, 184, 189, 191, 201, 203, 206, 209, 214, 222, 228, 231, 236, 251, 252, 256, 258, 262, 265, 268, 271, 272, 273, 274, 281, 284, 290, 299, 300, 301, 302, 306, 311, 314, 317, 330, 332

C Visualizations for different techniques

Figure 39: SYSU-IR, Single-modal network visible light images: Examples for query images and corresponding gallery images with lowest distance

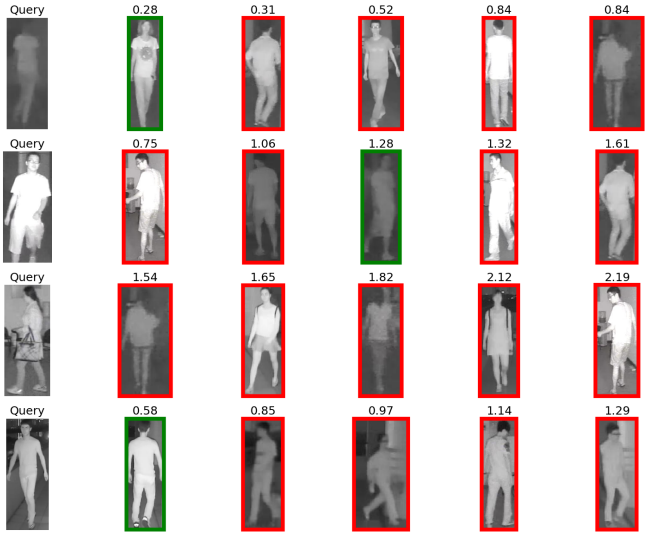


Figure 40: SYSU-IR, Single-modal network infrared images: Examples for query images and corresponding gallery images with lowest distance.

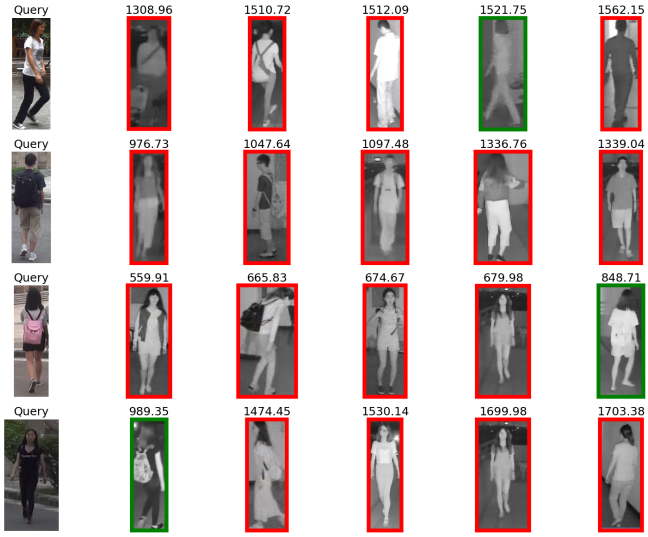


Figure 41: SYSU-IR, One-stream network: Examples for query images (RGB) and corresponding gallery images (infrared) with lowest distance.

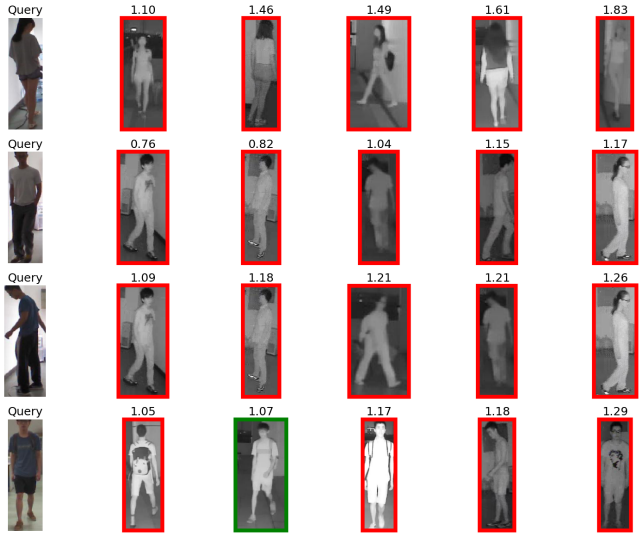


Figure 42: SYSU-IR, Cross-modal distillation network from infrared to RGB (triplet loss): Examples for query images (RGB) and corresponding gallery images (infrared) with lowest distance.

Figure 43: BIWI, Single-modal network visible light images: Examples for query images and corresponding gallery images with lowest distance

Figure 44: BIWI, Single-modal network depth images: Examples for query images and corresponding gallery images with lowest distance.

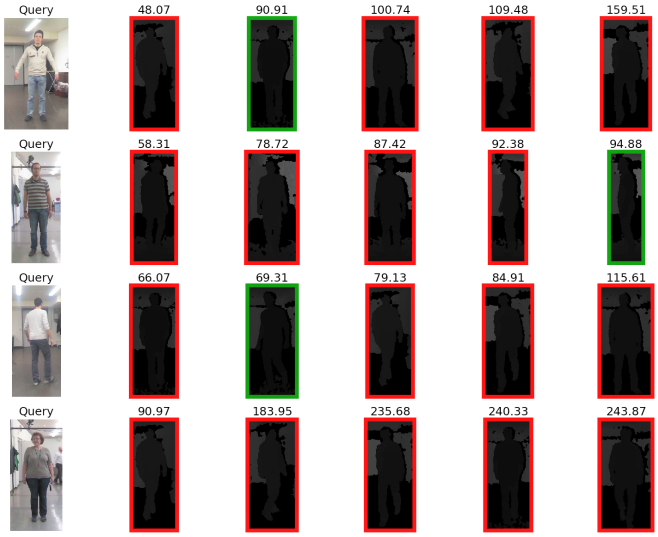


Figure 45: BIWI, One-stream network: Examples for query images (RGB) and corresponding gallery images (depth) with lowest distance.

Figure 46: BIWI, Cross-modal distillation network from depth to RGB (softmax loss): Examples for query images (RGB) and corresponding gallery images (depth) with lowest distance.

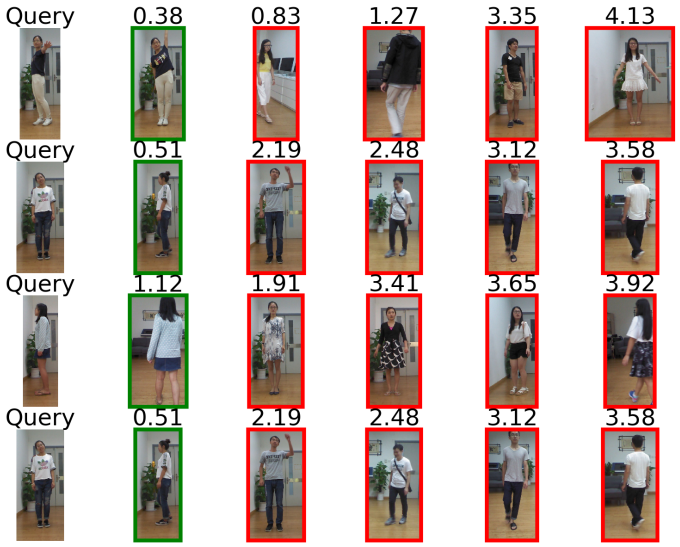


Figure 47: RobotPKU, Single-modal network visible light images: Examples for query images and corresponding gallery images with lowest distance

Figure 48: RobotPKU, Single-modal network depth images: Examples for query images and corresponding gallery images with lowest distance.

Figure 49: RobotPKU, One-stream network: Examples for query images (RGB) and corresponding gallery images (depth) with lowest distance.

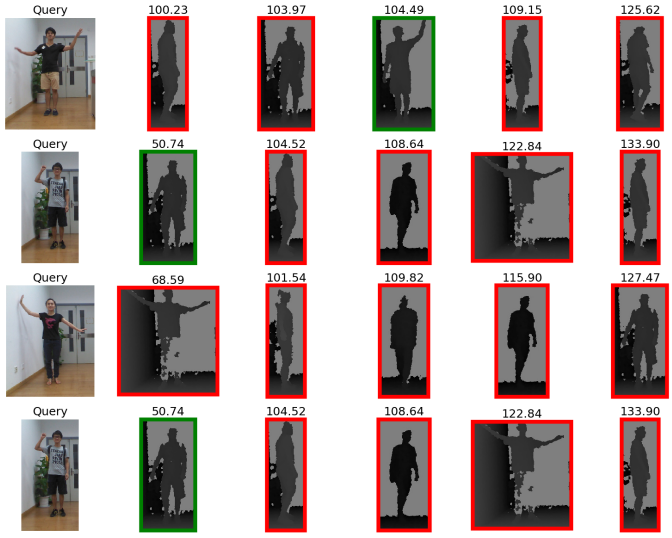


Figure 50: RobotPKU, Cross-modal distillation network from depth to RGB (softmax loss): Examples for query images (RGB) and corresponding gallery images (depth) with lowest distance.

D Paper "A Cross-Modal Distillation Network for Person Re-identification in RGB-Depth"

A Cross-Modal Distillation Network for Person Re-identification in RGB-Depth

Frank Hafner, Amran Bhuiyan, Julian F. P. Kooij, Eric Granger, *Member, IEEE*

Abstract—Person re-identification involves the recognition over time of individuals captured using multiple distributed sensors. With the advent of powerful deep learning methods able to learn discriminant representations for visual recognition, cross-modal person re-identification based on different sensor modalities has become viable in many challenging applications in, e.g., autonomous driving, robotics and video surveillance. Although some methods have been proposed for re-identification between infrared and RGB images, few address depth and RGB images. In addition to the challenges for each modality associated with occlusion, clutter, misalignment, and variations in pose and illumination, there is a considerable shift across modalities since data from RGB and depth images are heterogeneous. In this paper, a new cross-modal distillation network is proposed for robust person re-identification between RGB and depth sensors. Using a two-step optimization process, the proposed method transfers supervision between modalities such that similar structural features are extracted from both RGB and depth modalities, yielding a discriminative mapping to a common feature space. Our experiments investigate the influence of the dimensionality of the embedding space, compares transfer learning from depth to RGB and vice versa, and compares against other state-of-the-art cross-modal re-identification methods. Results obtained with BIWI and RobotPKU datasets indicate that the proposed method can successfully transfer descriptive structural features from the depth modality to the RGB modality. It can significantly outperform state-of-the-art conventional methods and deep neural networks for cross-modal sensing between RGB and depth, with no impact on computational complexity.

Index Terms—Deep Learning, Convolutional Neural Networks, Transfer Learning, Distillation Networks, RGB-D Vision, Person Re-Identification, Autonomous Driving, Video Surveillance.

1 Introduction

Person re-identification is an important function in many monitoring and surveillance applications, such as multi-camera target tracking, pedestrian detection in autonomous driving, access control in biometrics, search and retrieval in video surveillance, and forensics [1], [2], [9], and, as such, has gained much attention in recent years. Given the query image of an individual captured over a network of distributed cameras, person re-identification seeks to recognize that individual based on a gallery of previously-captured images [3].

Traditionally, person re-identification involves recognizing individuals over a network of non-overlapping cameras that sense in the same RGB modality. State-of-the-art single modal methods based on RGB images can be categorized as either feature learning based methods, that seek to learn robust and discriminant feature representations from person samples [4], [20], [21], [25], [26], [27], or distance learning based methods, that seek to learn an effective distance metric that can minimize the difference between persons from different cameras [5], [24], [28], [29], [30]. Single-modal re-identification remains a very challenging problem due to low resolution images, occlusions, misalignments, background clutter, motion blur, and variations in pose and illumination. Moreover, most of the state-of-the-art methods [20], [21], [22], [23], [24], [56] rely on the assumption that people do not usually change their clothing, i.e., their appearance across views remains same, which is unsuitable for long-term monitoring and surveillance.

New sensors to capture high-definition signals, like lidars and radars which sense in the depth modality, allow to expand on sensing capabilities, and are paving the way for innovative, next-generation monitoring and surveillance technologies. This paper focuses on deep neural networks for cross-modal person re-identification that allow sensing between RGB and depth modalities. Deep neural networks are highly successful at performing high-level visual recognition tasks due to their capacity to learn important low- and intermediate-level features from the raw image data. These networks are trained with labeled image data from both modalities, and then allow to recognize a person captured using either the RGB or depth sensor. Note that these networks differ from methods in literature for multi-modal person re-identification, where RGB and depth representations are combined (often normalized and concatenated) to improve performance [16], [74], [75], [76], [77].

Although some methods have been proposed for cross-modal re-identification between RGB and infrared images [10], [11], [12], [13], few address RGB and depth images [17]. However, sensing across RGB and depth modalities is important in many real-world scenarios. This is the case, for example, with video surveillance systems that must recognize individuals in poorly illuminated environments. Recent progress in lidar technology makes the usage of depth information more and more viable in these situations as a replacement for infrared cameras [14], [15]. Another example is the case of autonomous self-driving vehicles, which require tracking pedestrians around their vicinity,

A. Bhuyian and E. Granger are with the Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), École de technologie supérieure, Université du Québec, Montreal, QC H3C 1K3, Canada. F. Hafner was kindly hosted in LIVIA for this research. (e-mail: amran.apece@gmail.com; eric.granger@etsmtl.ca).
F. Hafner and J. Kooij are with the Intelligent Vehicles Group, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands (e-mail: f.m.hafner@student.tudelft.nl, j.f.p.kooij@tudelft.nl)

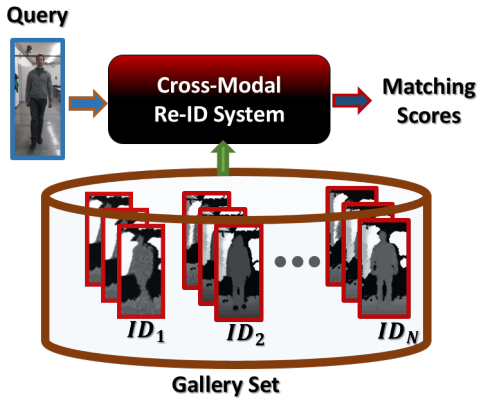


Fig. 1. Illustration of the cross-modal person re-identification system based on RGB (query) and depth (gallery set) modalities.

where some regions are covered by lidars sensors, and others by RGB cameras.

In this case, figure 1 illustrates an example of a cross-modal system. Query images captured for a person sensed in the RGB modality (captures color intensities), are matched against a set of reference gallery images from a depth modality (captures 3D geometry). There is a considerable shift across modalities since data from RGB and depth images are heterogeneous. State-of-the-art methods proposed for cross-modal re-identification are typically optimized co-jointly using image data from both source and target modalities which leads to complex re-identification models.

Cross-modal re-identification can be seen as a transfer learning task [31], [32], commonly employed to adapt visual recognition models to operate across domains in, e.g., image classification [33], [34], human activity classification [35], [36] and objection recognition [8], [37], [38], [39]. In particular, labeled data is available in both source and target domains. A same task (re-identification) is associated with two different domains, where source and target distributions differ [7]. The source domain corresponds to either the depth or RGB modality, while the target domain corresponds to the other modality. The main objective is to transfer the knowledge learned from source to target domain, even though the data distribution between the domains can incorporate a significant shift. Then, the cross-modal re-identification can recognize across two domains and, therefore, solve the transfer learning task (re-identification based on either RGB or depth) in a common representation space.

A key challenge for designing cross-modal networks is training when image data distribution incorporate a significant shift. State-of-the-art deep learning methods will typically optimize using image data from both source and target modalities jointly. Low accuracies in comparison to single-modal re-identification within infrared-RGB cross-modal re-identification suggest that co-joint optimization might not be ideal with respect to the significant distribution shift between the modalities [10], [11], [12], [13]. Hence, the technique proposed in this paper seeks to exploit the asymmetrical relationship between depth and RGB images by training the re-identification system in a sequential manner.

In this paper, a new cross-modal distillation network is proposed for robust person re-identification across RGB and depth sensors. Inspired by the unsupervised distillation method of [8], this paper adopts a deep neural network able to transfer learned representations from one sensor modality to another. The proposed approach relies on paired labeled images from both modalities for training, but is independent of paired images during testing or inference. Using a two-step optimization process, the proposed method transfers supervision between modalities such that similar structural features are extracted from both RGB and depth modalities. Extracting these features yields a discriminative mapping to a common feature space. A research goal of this work is to justify the ideal order of transfer, i.e., which modality is source and which one is target [39]. In the first step a network is optimized based on data from the first (source) modality, and then, in the second step, the embeddings and weights of this first neural network provide guidance to optimize a second network for the other (target) modality. Following [8], this cross-modal distillation network is initialized with the weights of the network trained in the first step, to facilitate the transfer of the knowledge to the other network. All mid-level to high-level layers of the second network are frozen during training. In contrast to [8], the optimization is based on the final embedding layer of the networks to guarantee an embedding in a common feature space for both modalities. Note that the proposed distillation network is a general model for cross-modal re-identification that may be extended to other combinations of modalities and to recognize other visual objects in image retrieval (e.g., vehicles) where appearance changes. However, in order to better understand the asymmetrical relationship between depth and RGB modalities, this paper focuses on recognizing persons across depth and RGB.

This paper presents the following contributions: *(i)* A deep cross-modal network is adopted to transfer an embedding representation from one modality to the other by exploiting the intrinsic relation between depth and RGB. *(ii)* In contrast to the majority of literature in person re-identification we investigate the choice for a certain embedding size and embedding layer with experiments. We are able to show, that an embedding extracted from the softmax classification layer can be competitive to the commonly used preliminary layer embedding. *(iii)* Extensive experimental validation is conducted to show the advantages of the proposed method over state-of-the-art networks on multiple RGB-D based benchmark re-identification datasets. To our knowledge, this is the first deep cross-modal distillation network for re-identification between RGB and depth.

The rest of the paper is organized as follows. Section 2 provides an overview of conventional, deep learning and cross-modal techniques related to person re-identification. Section 3 describes deep cross-modal neural network techniques as well as the proposed cross-modal distillation network. Section 4 describes the experimental methodology (dataset, protocol and performance metrics) used for validation of the proposed and baseline systems, and section 5 presents the experimental results. Finally, Section 6 describes our main findings, and highlight directions for future research.

2 Related Work

The area of person re-identification has received much attention in recent years [9]. This section provides a summary of the state-of-the-art conventional, deep learning and cross-modal techniques as they relate to our research.

Conventional Methods. Conventional approaches for person re-identification from a single modality can be categorized into two main groups – direct methods (with handcrafted descriptors or learned features) and metric learning based approach. Direct methods for re-identification are mainly devoted to the search of the most discriminant features, or combinations thereof, to design a powerful descriptor (or signature) for each individual regardless of the scene. In contrast, in metric learning methods, a dataset of different labeled individuals is used to jointly learn the features and the metric space to compare them, in order to guarantee a high re-identification rate.

Due to the non-rigid structure of the human body, it is difficult to model the appearance of the whole body for re-identification. Instead it is more robust to model the appearance focusing on salient parts or meaningful parts of the body. Most of the direct method based re-identification approaches rely on the local meaningful parts, e.g. horizontal stripes [24], [42], triangular graphs, concentric rings [43], symmetry-driven structures [20], pictorial structure [22], meaningful body-parts [21] and horizontal patches [44]. Different features (such as: Color based features [20], [21], [22], textures [45], [46], [47], edges [47], Haar-like features [48], interest points [49] and Biologically Inspired Features (BIF) [49]) and different combination of those features (such as: Bag-of-Words (BoW) [78], Weighted Histogram of Overlapping Strips (WHOS) [73], & Local Maximal Occurrence (LOMO) [24]) from those local regions have proven to be useful to achieve better re-identification accuracy. Given the handcrafted features, another stream of direct method based re-identification approaches learns the feature importance based on the salient feature analysis of each individual [4], [21], [27], or by exploiting the coherence among different features on manifold space [72].

Metric learning based approaches usually find a mapping from feature space to a new space in which feature vectors from image pairs of the same individual are closer than feature vectors from different image pairs. Commonly used metric learning techniques that are adopted for re-identification include Mahalanobis metric learning [53], Large Margin Nearest Neighbor Learning (LMNN) [52], Logistic Discriminant Metric Learning (LDML) [52], Kernel Canonical Correlation Analysis (KCCA) [54], keep it simple and straight forward metric learning (KISSME) [53] and Cross-view Quadratic Discriminant Analysis (XQDA) [24].

Deep Learning Methods. Similar to other vision applications, there has also a been growing number of deep learning based re-identification approaches [55], [56], [57], [58], [59], [60], [61], [62], [63]. The idea of using a deep learning architecture for person re-identification stems from Siamese CNN with either two or three branches for pairwise verification loss [55], [56], [57], [58], [59], [62] or triplet loss [60], [61], [63] respectively, or combination of both [64]. Some of those approaches use their own network architectures, by proposing new layers [56] or by fusing features from different body parts with a multi-scale CNN structure [57], [65]. Some other [60], [63], [68] use the pre-trained or different variants of pre-trained models (e.g. Resnet [41], GoogleNet [66]) which often obtain great re-identification accuracy. Another trend of using deep learning architecture is *transfer learning* [59], [70], [71], for when the distribution of the training data from the source domain is different from that of the target domain. The most common deep transfer learning strategy for re-identification [70] is to pre-train a base network on a large scale source dataset, and transfer learned representation to the target dataset. Variant of other transfer learning approaches for re-identification [59], [71] leverages the idea of joint or multi-task learning considering combination of different re-identification datasets, or auxiliary datasets to minimize the cross-domain discrepancy. However, these transfer learning methods depend on the assumption that the tasks are the same and in a single modality. Thus all the above mentioned single modality based approaches are unsuitable when the source and target domains are heterogeneous.

Cross-Modal Methods. While the progress in re-identifying persons in single modalities was significant, only few works [10], [11], [12], [13], [16], [17] investigated the task of cross-modal person re-identification.

Recently, several works were published concerning cross-modal person re-identification between RGB and infrared images [10], [11], [12], [13]. In [10], the authors analyze several standard neural network structures to embed the RGB vs. IR modalities in a common feature space on their proposed SYSU-IR dataset. The key architectural contribution is the 'One stream structure with zero-padding augmentation' network. The zero-padding network as well as the simple one-stream network from [10] will be analyzed in more detail in section 3.2. In [11], the authors presented a two-stream neural network which combined a contrastive and a softmax loss together. To enhance the results they attached a subsequent metric learning step. A similar scenario was also used in [12] where the authors adopt the same methodology as [11] and combine two losses. The first loss has the goal to minimize the cross-modal intra-distance and at the same time maximize the inter-modal distances. Hence, the authors compare the distance of a positive visible-thermal image pair and the minimum distance of all negative visible-thermal pairs. This loss, is accompanied by an identity loss to guarantee the robustness. The cross-modal re-identification problem on RGB-IR scenario has been addressed in adversarial way in [13]. The idea of the authors is to combine three losses. The first two losses are a identity loss and a triplet loss. Additionally, they introduce a GAN based structure on their network architecture. The discriminator differentiates from which modality the input sample came and, hence the generator enforces a mutual embedding.

There are a few works in the literature that consider *multi-modal* person re-identification scenario [16], [74], [75], [76], [77] by fusing the RGB and the depth information in order to extract robust discriminative features. In [74], the authors fused clothing appearance features with anthropometric measures extracted from depth data. In [75], a tri-modal based person re-identification method has been proposed by combining the RGB, depth and

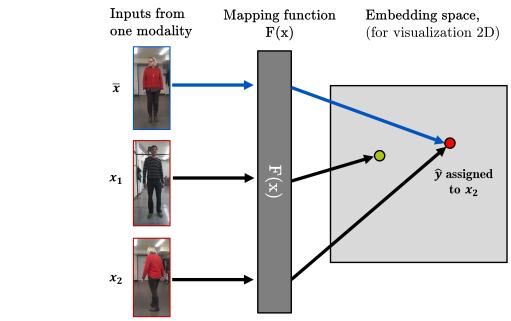
thermal data. In [76], the authors proposed a height-based gait feature that integrated RGB based height histogram and gait feature from depth data. In [77], a depth based segmentation technique is used to extract the features from the foreground body parts. In [16], a depth-shape descriptor called eigen-depth is proposed to extract describing features from the depth domain. The distance between eigen-depth features are proven to lie in Euclidean space and are rotation invariant. The authors were able to show that those orientation-invariant descriptors of body regions are less prone to errors from position and lighting changes. Additionally, the authors defined a common latent subspace for the eigen-depth features and features extracted in the RGB modality. Although, the methodology is in principle applicable in cross-modal re-identification, the authors did not perform any evaluations in this domain [16]. Finally [17] used the same features to perform cross-modal re-identification between depth and RGB.

In 2016, Gupta et al. [8] presented a transfer learning network for cross-modal distillation. Their goal is to use learned representations from large datasets in a certain modality for classification in a paired modality with limited labeled data. An example usage of this network is the transfer of the capabilities of a CNN object classifier in RGB to the corresponding depth images. Therefore, the network trained in the RGB modality is copied to the depth modality. Afterwards a mid-level layer in the network is frozen and optimized by means of unlabeled coupled images. Hence, a common mid-level layer is enforced, while the low-level features can be learned in the new modality.

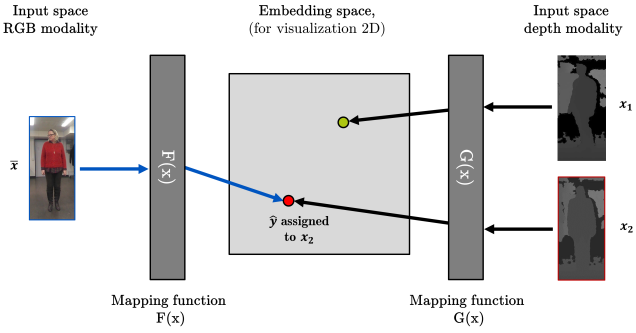
In contrast to the above works on cross-modal re-identification, we propose to employ the cross-modal distillation idea by means of a deep transfer learning technique. The idea of the method is inspired by the recent work on supervision transfer of Gupta et al. [8]. However, supervision transfer [8] and our approach aim at different problems with different focuses of method design: supervision transfer solves the problem of limited data availability for object detection problems with a transfer scheme from RGB to depth. Our method is using the distillation paradigm to transfer knowledge from one modality to a second modality to solve the re-identification task across the two modalities. Therefore, contrary to Gupta et al. [8], the task has to be solved across modalities in the same feature space and is not considered a pre-training procedure as in [8]. In Gupta et al. the direction of transfer is defined as from RGB to depth. In contrast, in this work the ideal direction of transfer is one of the research questions which is answered.

3 Deep Cross-Modal Neural Networks

In this section deep neural networks are presented for cross-modal person re-identification based on RGB and depth modalities. These networks are trained with labeled image data from both modalities. During inference, the trained network then allows to recognize the same person captured using either the RGB or depth sensor. To date, no deep neural networks architectures have been applied to solve the cross-modal person re-identification between RGB and depth.



(a) Single-modal re-identification



(b) Cross-modal re-identification

Fig. 2. (a) Single-modal re-identification embeds input (from the same modality) to a common latent feature space, such that different images from the same individual are close together in the mapping. (b) Cross-modal re-identification creates a shared embedding for multiple modalities, each with their own mapping function.

Consider a query image $\hat{x}$, and a set of gallery images $x_1, \cdots, x_M$ with associated labels $y_1, \cdots, y_M$, such that y_i indicates the individual present in image x_i. In single-modal re-identification, both query $\hat{x} \in \chi$ and gallery images $x_i \in \chi$ are from the same input space χ. The general approach to person re-identification is to apply a mapping from the input images to an embedded space, where input samples of the same individual are mapped close together, and of different individuals are further apart. Figure 2a shows how this embedding is used during test time for the standard single-modal case with RGB colour images. The query image $\hat{x}$ is mapped to the embedded space $F(\hat{x})$, where the distances to the gallery images $F(x_i)$ are compared. The identified person $\hat{y}$ for query $\hat{x}$ is then the individual corresponding to the closest embedded gallery image $\hat{i}$, i.e.

$$\hat{y} = y_{\hat{i}} \qquad \text{where} \qquad \hat{i} = \underset{i}{\operatorname{argmin}}\ d(F(\hat{x}), F(x_i)). \tag{1}$$

where d is the distance metric for the embedding, typically the Euclidean distance $d(a, b) = \|a - b\|$. During training, the learning objective is therefore to estimate a suitable mapping $F(x)$ from available training data.

For cross-modal re-identification an additional challenge is added, as query and gallery images can now use different input spaces. Figure 2b shows an example with a depth image as query, using RGB gallery images. Since both input spaces now have to be mapped to the same latent space, hence training involves the additional challenge of learning

a mapping $G(x)$ for depth images to the shared feature space with $F(x)$.

In our work, the cross-modal re-identification task is formulated as a transfer learning problem, where labeled data is available in both source and target domains. D^s is defined as the source domain, while D^t is the target domain. In the case of cross-modal sensing between RGB and depth, D^s corresponds to either the depth or RGB modality, and D^t corresponds to the other modality. A domain D consists of an input space χ with a marginal probability distribution $P(\chi)$. In our case, there is a considerable shift across domain distributions, since RGB and depth images are heterogeneous, and thus $\chi^s \neq \chi^t$. A task T is defined by a label space, and in our case, both modalities are related to the same person re-identification task. The task in the source domain is denoted as T^s, while the task in the target domain by T^t. Hence, cross-modal person re-identification can be seen as a case of transfer learning where a shared re-identification task $T^s = T^t$ is associated with two different domains $D^s \neq D^t$, where either the source and target data representations or the source and target distributions differ [7]. Additionally, the cross-modal re-identification seeks to recognize across two domains and, therefore, solve the tasks T^s and T^t in a common feature space, instead of each task separately.

To formalize our approach, section 3.1 will first present common deep neural network architectures and loss functions which were successful applied for single-modal re-identification. Then, using these components, section 3.2 first presents two cross-modal baseline approaches taken from existing work on person re-identification between RGB and infrared. Section 3.3 will then introduce our main contribution, the cross-modal distillation network for RGB and depth.

3.1 Methods for Single-Modal Re-Identification

In most research on person re-identification, both modalities are the same, $D^s = D^t$. Therefore, the task is defined as a single-modal re-identification problem. For this task, several successful feature extraction networks and loss functions have been employed to train deep learning architectures for person re-identification. Although, we cannot cover all feature extractors and losses in this paper, this section presents common ones which were successful applied for single-modal re-identification.

For feature extraction, our work uses Residual neural networks (Resnet) [41] which are pre-trained on ImageNet. The Resnet architectures were shown to be effective for several person re-identification applications [67], [69] . The general Resnet architecture consists of convolutional blocks with residual connections to enable learning in deep networks. To assess the influence of a shallow Resnet network versus a deeper one, both Resnet18 and Resnet50 are explored. Furthermore, we consider two possible loss functions, triplet loss and softmax loss, which both have been successfully applied in single-modal person re-identification [9], [60], [68]. These losses are used to learn embeddings for the input images, such that images of the same individual have a small Euclidean distance in the embedded space, while distinct individuals are far apart. We will now shortly discuss both losses in more detail.

3.1.1 Triplet Loss

Using the *triplet loss* results in a metric learning approach which directly optimizes an embedding layer in a certain distance metric. During training, this loss compares the relative distances of three training samples, namely a so-called anchor image x_a, a positive image sample x_p from the same individual as x_a, and a negative sample x_n from a different individual. Given an anchor image x_a, this loss assures that the embedding of an image taken from the same class x_p is closer to the anchor's embedding than that of a negative image belonging to another class y_n by at least a margin m in distance metric d. In the following, F denotes the deep neural network structure to optimize, correspondingly $F(x)$ is the result of a forward pass with image x through the network to the final embedding layer. Anchor image x_a and positive image x_p are extracted from an instance with the same label $y_a = y_p$. The negative image is defined as x_n and is taken from another instance, hence $y_a \neq y_n$. The triplet loss is therefore defined as

$$L_{tri} = \sum_{i=1}^{T} \left[d(F(x_{a(i)}), F(x_{p(i)})) - d(F(x_{a(i)}), F(x_{n(i)})) + m \right]. \tag{2}$$

Here, indices $a(i)$, $p(i)$ and $n(i)$ stand for anchor, positive and negative, of the i-th triplet, and T for the number of triplets used per batch.

3.1.2 Softmax Loss

For the second considered loss, the *softmax loss*, the embedding is learned indirectly by first treating re-identification on the training set as a classification problem, where all C individuals in the training set are considered a different class. During training, the softmax loss thus optimizes the class probabilities for the instances in the training set. Afterwards, a layer of the neural network prior to the softmax loss is used as the embedding. This enables that the network can be applied on test data, which can contain new individuals not present in the training data, by only keeping the network output $F(x_i)$ at a layer before the softmax function, which is considered the M-dimensional embedding for test images x_i. In literature for re-identification the embedding layer is usually chosen as the penultimate layer before the softmax loss [9]. Therefore, the softmax loss to optimize the embedding can be expressed as

$$L_{soft1} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{e^{W_{(y_i)} F(x_i) + b}}{\sum_{j=1}^{C} e^{W_{(j)} F(x_i) + b}} \right), \tag{3}$$

where N is the batch size, $W_{(j)}$ are the weights leading to the j-th node of the ultimate softmax layer of the network, b is a bias and M is the variable amount of nodes in the penultimate layer. The amount of classes is defined as C.

Apart from the common embedding $F(x)$, our work also investigates including the final transformation $F'(x) = WF(x) + b$ as an alternative C-dimensional softmax embedding. Note that the embedding size is now fixed to the amount of classes C in the training set.

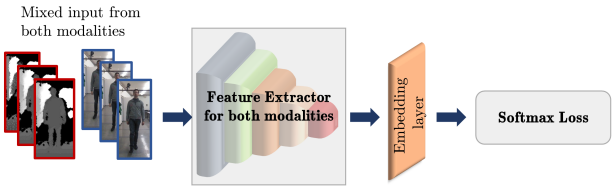


Fig. 3. Cross-modal architectures based on a one-stream network.

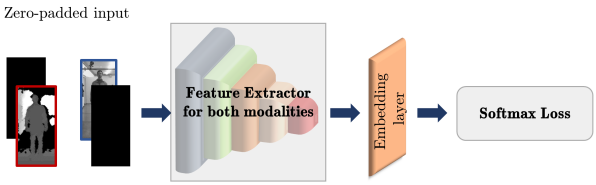


Fig. 4. Cross-modal architectures based on a zero-padding network.

Using $F'_{(j)}(x_i)$ to denote the j-th element in this C-dimensional embedded vector $F'_{(j)}$, the softmax loss for this alternative embedding can now be written as

$$L_{soft2} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{F'_{(y_i)}(x_i)}}{\sum_{j=1}^{C} e^{F'_{(j)}(x_i)}}\right). \quad (4)$$

3.2 Cross-Modal Architectures for Re-Identification

We now introduce two state-of-the-art cross-modal networks from the literature on re-identification across RGB and infrared, which we will apply to re-identification across RGB and depth. Both these methods are optimized co-jointly using image data from RGB and depth modalities. They approach tasks T^s and T^t for cross-modal re-identification in a parallel manner, since images from both modalities are provided to the network in mixed batches. Therefore, in these cases the mapping functions are identical, $F(x) = G(x)$.

The first cross-modal architecture is the *one-stream neural network*, which is illustrated in figure 3. It is designed in the same way as a CNN for single-modal re-identification, using a Resnet feature extractor and softmax loss [10]. The only difference for optimization as explained in section 3.1 is that the weights are optimized with mixed batches of both modalities. These images are provided equally to the network and, therefore, no outer guidance concerning modality-specific nodes in the network is given.

The second cross-modal architecture, the *zero-padding neural network* from [10], is shown in figure 4. It incorporates two input channels, and the key idea is to embed each modality in a separate channel and pad the other channel with zeros. By using the zero-padding of one channel in each modality, several nodes in early layers within the network are influenced by only one of the two modalities. Therefore, the network obtains outer guidance on specific nodes for the first modality, specific nodes for the second modality and shared nodes. This architecture is also based on Resnet feature extractor and optimized using softmax loss.

3.3 A Cross-Modal Distillation Network

This subsection introduces our novel cross-modal approach. The major difference to the approaches presented in the previous subsection is that the tasks T^s and T^t are approached in a sequential manner, rather than in parallel. Therefore, the training of the task in the source modality is separated from the training of the task in the target modality. The conceptual cross-modal distillation scheme to transfer the supervision from one modality to the other modality is adapted from the work by Gupta et al. [8], see section 2. Nevertheless, several crucial differences to the cross-modal distillation of Gupta et al. are existent which were elaborated in section 2. The main objective of the sequential cross-modal distillation is to exploit the intrinsic relation of the two modalities to be able to extract similar features from both. The training of the network is divided into two steps, as visualized in figure 5, which will be explained in detail next.

3.3.1 *Step I – Training of the Baseline Network*

In step I of the training of the cross-modal distillation network, a neural network F is trained for sensing in a first modality D^s, as presented in section 3.1. The feature extractors Resnet18 and Resnet50 as well as softmax loss and triplet loss will be used to optimize networks for the baseline of the cross-modal distillation network (for more details see chapter 3.1). The network is optimized by means of an early-stopping criteria based on the mAP in the validation set. Afterwards, the network is frozen as F_{fr}, with corresponding weights $W_{F,fr.}$.

3.3.2 *Step II – Cross-Modal Distillation*

The obtained neural network feature extractor for the first modality is deployed as the baseline network for the training of a feature extractor for the second modality. For the second training step, a network with the same architecture as the corresponding network in step I is initialized.

Similarly to [8], the weights of the converged model from step I, $W_{F,fr.}$, are copied to network G which is dedicated to the second modality. Additionally, the weights of the network are frozen from a mid-level convolutional layer up to the final feature embedding. This retains the high-level mapping from the first network, which was successfully trained in the source modality, to the target modality. At the same time, the target embedding can still learn meaningful low-level features for the task in the target modality.

For the actual transfer of knowledge we make use of paired images X_{m1} from modality 1 and X_{m2} from modality 2. The aim is to optimize G in such a way that the embeddings of images from the second modality X_{m2} with label y are close to the embeddings of images from the first modality X_{m1} with label y. This is realized by exploiting image pairs $x_{m1,i}$ and $x_{m2,i}$ from the two modalities, which are considered coupled as they are taken at the exactly same time step. Hence, the embedding of $x_{m1,i}$ is obtained with a forward propagation through the frozen network $F_{fr.}$ and is taken as the groundtruth for the embedding of $x_{m2,i}$ with the, at this stage, trainable network G. Since during inference mode the embeddings will be compared based on Euclidean distance, we aim to minimize this metric between

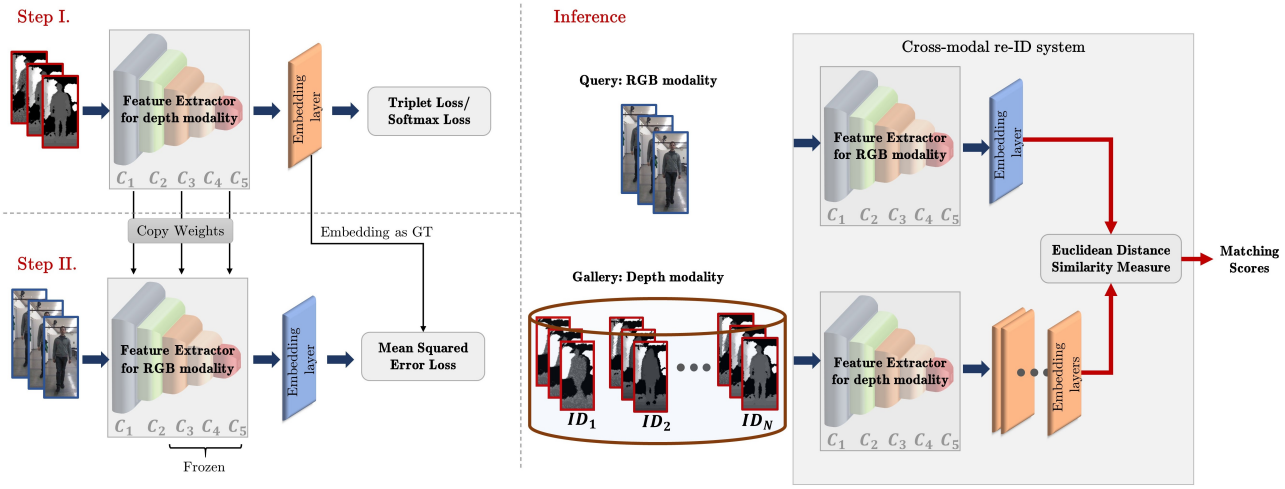


Fig. 5. Two step training scheme and inference for the proposed cross-modal distillation network. Step I involves training of a CNN for single-modal re-identification. In step II, the knowledge from the first modality is transferred to the second one. During inference, query and gallery images different modalities produce feature embeddings and matching scores for cross-modal re-identification. As an example, this figure is exemplary of a transfer from depth to RGB, and a inference with RGB as query and depth as gallery. The modalities can be interchanged in both cases.

Algorithm 1 Cross-Modal Distillation Network

Input: Input Train Data with paired images, X_{m1}, X_{m2}

STEP I: Training baseline network

$j = 0$
$mAP_{val,best} = 0$
Initialize network F with parameters W_F using a pre-trained CNN
while $(j < MAXEPOCH)$ **do**
  Perform training of F, train (X_{m1},W_F) using loss function 2 or 3.
  if $mAP_{val,j} > mAP_{val,best}$ **then**
    save W_F as $W_{F,best}$
  end if
  $j = j + 1$
end while

STEP II: Cross-modal distillation

$j = 0$
$L_{val,best} = \infty$
Load $W_{F,best}$ into F and freeze to $F_{fr.}$
Initialize weights W_G of network G with weights $W_{F,best}$
Freeze mid- to high-level weights of W_G
while $(j < MAXEPOCH)$ **do**
  Perform training of G, train (X_{m2}, W_G) using loss function 5 and $F_{fr.}(X_{m1})$ as groundtruth
  if $L_{val,j} < L_{val,best}$ **then**
    save W_G as $W_{G,best}$
  end if
  $j = j + 1$
end while
Load $W_{G,best}$ into G and freeze to $G_{fr.}$

Output: Models $F_{fr.}$ and $G_{fr.}$

the two embeddings. Hence, we make use of the mean squared error (MSE) loss between the embeddings of paired images $F_{fr.}(x_{m1,i})$ and $G(x_i)$ which is defined as

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|F_{fr.}(x_{m1,i}) - G(x_i)\|^2 \quad (5)$$

where N is the batch size in training stage. The weights W_G of network G are optimized based on this loss function and trained until convergence. Early-stopping criteria for the training of this network is the loss in the validation set. The whole training procedure is formalized algorithm 1.

3.3.3 Inference

In inference mode, the two resulting neural networks $F_{fr.}$ and $G_{fr.}$ are evaluated in the corresponding modalities to obtain feature embeddings for input images. Similarity between the feature representations is measured using Euclidean distance. For each query image, each gallery image is therefore ranked according to the similarity between embeddings in Euclidean space, and the label of the most similar gallery image is returned, see equation (1).

4 Experimental Methodology

In this section we present the experimental methodology used to validate the proposed approach. Therefore, two RGB-D person re-identification datasets will be presented. As these datasets were originally not designed for cross-modal person re-identification it is important to discuss their intrinsic properties and the adjustments in detail. Additionally, a complete description of the evaluation protocol used in this work will be given to enable repetition of the experiments.

4.1 Datasets

Two publicly-available dataset for person re-identification were considered for the experiments, namely BIWI RGBD-ID [18] and RobotPKU [19] datasets. These datasets were selected because they provide high-resolution depth and RGB images, a decent amount of instances and a large amount of images per instance in different poses. These are prerequisites to successfully train neural networks for re-identification. No other public datasets which were found were satisfying these requirements.

The BIWI RGBD-ID dataset targets long-term people re-identification from RGB-D cameras [18]. The dataset is recorded with a Microsoft Kinect, which provides depth, RGB images and a skeleton. The skeleton is neglected for this work. As in [17] same person with different clothing is considered as a separate instance. Overall, it is comprised of 78 individuals with 22,038 images in depth and RGB. The BIWI dataset consists of RGB images with a resolution of 1280×960 and depth images with a resolution of 640×480. In all images the individuals were cropped out in RGB and depth with a margin in all directions and resized to 256×128 for training. RGB and depth images are provided coupled with no visible difference in capturing time.

As with the BIWI dataset, the RobotPKU dataset was captured with a Microsoft Kinect camera [19]. The dataset consists of 90 persons with 16,512 images in total. The depth and RGB images in the RobotPKU dataset are provided cropped, and hence, the images have varying resolutions corresponding of the distance of the individual to the camera. For training, all images are resized to 256×128. The images are provided in a coupled manner. Nevertheless, by visual inspection it is apparent, that there is a slight time difference, in the order of a fraction of a second, between the images captured in depth and RGB. Compared to the BIWI dataset, the depth images in the RobotPKU dataset are more noisy and often body parts, like heads and arms, are absent in the images.

Although RGB-infrared re-identification within the SYSU-IR dataset [10] is considered a parallel stream to RGB-depth re-identification no evaluations on this dataset will be made. This is due to the fact, that the cross-modal distillation network is primarily designed for the properties of RGB and depth [8]. Additionally, in this dataset no paired images of the modalities are available.

4.2 Evaluation Protocol

For the performance evaluation with the BIWI dataset, the same partitions into training, validation and testing subsets were adopted as in [17]. Accordingly, the dataset is divided into videos from 32 individuals for training, 8 instances for validation and 38 individuals for testing. For the RobotPKU dataset, the division will be videos from 40 individuals for training, 10 for validation, and 40 for testing. This follows the division of [19]. The exact split (label of individuals used to form subsets) is provided in appendix A.

For quantitative evaluation, the average rank 1, 5 and 10 accuracy performance measure is reported along with the mean average precision (mAP). For the reporting of the rank accuracy, a single-gallery shot setting is used, where a random selection of the gallery (G) images is repeated 10 times. For the query (Q) a maximum of 50 images per person are randomly selected. For the evaluation of cross-modal performance images of all cameras are compared. The only exception to this is the removal of the exactly same corresponding image in the parallel modality.

To obtain statistically reliable results for the proposed and baseline methods based on deep neural networks, average results are obtained through a 3-fold cross-validation process. The methods are trained and evaluated 3 times, and for each replication, a different validation subset is randomly extracted from within the design subset. Hence, the average values for performance measure are reported with standard deviation.

5 Experimental Results

An extensive series of experiments has been considered to validate the proposed cross-modal distillation network. In this section, the results for optimization with the single modalities (i.e., step I. in Fig 5) are first shown to establish a baseline for the individual modalities. Hence, we first investigate how different choices for deep networks and losses affect the performance on single-modal re-identification, and compare the relative difficulty of the modalities and dataset. Then, the distillation step (step II.) of the proposed method is performed and evaluated (section 5.2). Here, insights in how the distillation network is ideally trained are given. This involves the choice of the correct baseline network as well as the direction of transfer in the distillation step. Additionally, a sensitivity analysis of the results for the cross-modal distillation is performed (section 5.3). Finally, the presented method (section 5.4) is compared to other baselines and the state-of-the-art of the cross-modal person re-identification task between RGB and depth are defined. The findings of this section are underlined with an analysis of the activations of the neural networks (section 5.5).

5.1 Single-Modal Re-identification Performance

For performance evaluation with individual modalities (RGB and depth separately), several neural network optimizations have been investigated. Results have been obtained on BIWI and RobotPKU datasets using two architectures for feature extraction. The shallower network, Resnet18, and a deeper network, Resnet50. Both architectures have been optimized with triplet loss, equation (2), and softmax loss, equation (3).

For triplet loss an embedding size of 128 and a training batch of 64 with 16 instances á 4 images was used. As triplets the most difficult combinations within the batches were chosen. These parameters were proposed by [60].

For the following sections the standard softmax loss definition, equation (3), will be used with an embedding size of 128. This embedding size corresponds to the embedding size of triplet loss to enable a fair comparison of the optimizations. A more detailed analysis of the influence of the embedding size will be discussed in section 5.3 where a comparison of the best performing methods with different embeddings will be made. For this also the novel softmax loss definition in equation (4) will be evaluated. Corresponding to triplet loss a batch size of 64 will be used.

TABLE 1
Average test set accuracy of the proposed method (Step I) for different modalities on BIWI dataset.

Modality	Feature Extractor	Loss	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
RGB	Resnet18	Triplet	93.68 ± 0.76	99.65 ± 0.35	99.96 ± 0.04	94.77 ± 0.83
		Softmax	93.32 ± 1.83	99.67 ± 0.24	99.93 ± 0.09	94.46 ± 1.55
	Resnet50	Triplet	92.14 ± 1.86	99.71 ± 0.24	99.95 ± 0.08	93.44 ± 1.46
		Softmax	**94.75 ± 0.74**	**99.75 ± 0.19**	**99.96 ± 0.03**	**95.68 ± 0.60**
Depth	Resnet18	**Triplet**	**61.28 ± 2.49**	**93.85 ± 1.05**	**99.44 ± 0.18**	**62.71 ± 2.37**
		Softmax	57.09 ± 0.79	88.96 ± 0.15	96.95 ± 0.20	58.38 ± 1.07
	Resnet50	Triplet	54.23 ± 1.75	91.48 ± 0.56	99.15 ± 0.18	55.31 ± 1.71
		Softmax	59.84 ± 0.66	90.54 ± 0.81	97.80 ± 0.19	61.44 ± 0.54

TABLE 2
Average test set accuracy of the proposed method (Step I) for different modalities on RobotPKU dataset.

Modality	Feature Extractor	Loss	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
RGB	Resnet18	Triplet	**90.53 ± 0.65**	**99.30 ± 0.17**	**99.46 ± 0.10**	**91.91 ± 0.64**
		Softmax	84.73 ± 0.47	98.00 ± 0.12	99.24 ± 0.14	86.86 ± 0.46
	Resnet50	Triplet	89.04 ± 3.91	99.17 ± 0.33	99.46 ± 0.10	90.63 ± 3.41
		Softmax	84.52 ± 0.24	97.91 ± 0.35	99.12 ± 0.23	87.11 ± 0.22
Depth	Resnet18	Triplet	n/a	n/a	n/a	n/a
		Softmax	39.17 ± 0.34	69.85 ± 0.63	82.58 ± 0.35	38.65 ± 0.44
	Resnet50	Triplet	n/a	n/a	n/a	n/a
		Softmax	**44.50 ± 1.02**	**75.83 ± 1.29**	**87.56 ± 0.87**	**44.50 ± 1.02**

Table 1 shows the average accuracy of the networks for single-modal re-identification for individual (RGB and depth) modalities on BIWI data. Results show that the networks optimized using RGB modality alone, can reach a high level of accuracy. The best model, (Resnet50 optimized with softmax loss) provides an average mAP of 95.68%. The performance of networks optimized with triplet loss and softmax loss lead to comparable performance. As expected, the overall accuracy for the networks optimized using depth modality alone is much lower compared to the accuracy achieved for the same task with RGB. The highest accuracy (mAP = 62.71%) is achieved using the Resnet18 network optimized with triplet loss.

Table 2 shows the average accuracy for single-modal re-identification for individual (RGB and depth) modalities on RobotPKU data. Again, the RGB modality allows to achieve high level of accuracy. For instance, using Resnet18 trained with triplet loss yields the highest level of accuracy (mAP of 91.91%). Models trained with softmax loss generally obtain a slightly lower accuracy. In the depth modality, the networks using Resnet50 with softmax loss achieve an average mAP of 44.50%. Networks trained with triplet loss did not converge to produce meaningful embedding layers. This is caused by the inherent complexity of the re-identification task in the depth images of the RobotPKU dataset. This complexity is also reported in the performance indicators for the networks optimized with softmax loss. Overall results indicate that, compared to the BIWI data, the re-identification task is more challenging with the RobotPKU data, especially in the depth modality. This is explained by the higher level of noise in RobotPKU images, as well higher variability in the objects orientations.

The difference in performance for sensing in RGB and depth in both datasets gives insights in the complexity of the individual tasks. Following the results for both datasets, it is comparably easy to solely sense in RGB as visual cues like color features can be exploited very effectively for the re-identification task. In depth, color features are not present and the features based on a persons shape are less descriptive and lead to a lower accuracy. Nevertheless, it was shown that also in depth descriptive features can be extracted. The performance of the models in depth in BIWI is significantly higher than in the RobotPKU dataset. The lower accuracy for RobotPKU suggest that it is much more challenging to sense in the depth modality in this dataset. Therefore, it is expected that the transfer of features in RobotPKU is more difficult than in the BIWI dataset.

5.2 Performance for Cross-Modal Distillation

The cross-modal distillation network introduced in section 3 involves two optimization steps. In the previous section 5.1 networks for single-modal re-identification were analyzed. These networks correspond to the training in Step I of the cross-modal distillation. In this section experiments are presented to gain insight on the step II (distillation), and, in particular, on the advantages of transferring knowledge based on the depth or RGB modality.

Figure 6 presents the average mAP accuracy of the cross-modal distillation networks trained on the BIWI dataset in the cross-modal tasks with varying population of query and gallery between RGB and depth. The top four networks train the baseline network in depth (step I.), and then transfer to RGB (step II.). The bottom four networks train the baseline network in RGB (step I.), and then transfer to depth (step II.). Results are shown for the two feature extractor architectures Resnet18 and Resnet50. Additionally,

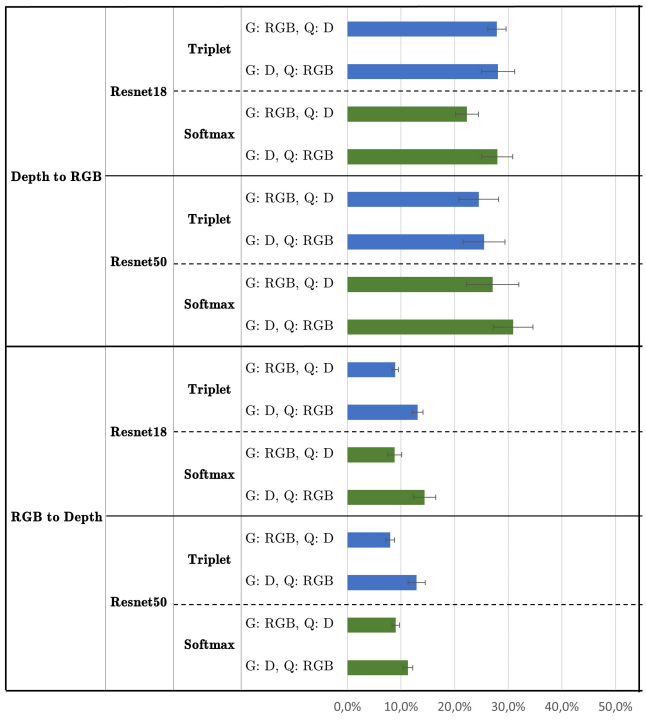


Fig. 6. Average mAP accuracy of various cross-modal distillation networks on the BIWI dataset. For all combinations we report varying query (Q) and gallery (G) modalities. The first column indicates the direction of the transfer for the cross-modal distillation.

the different colors indicate results with triplet (blue) and softmax (green) loss functions.

Results indicate that the accuracy obtained for when transferring from RGB to depth are significantly lower than from depth to RGB. Using depth images to populate a reference gallery, and RGB images as query achieves an mAP accuracy of about 31% using Resnet50 optimized with softmax loss. The best mAP accuracy for the same task and transferring from RGB to depth is about 13%. An explanation for this behavior is that the general shape information of a person that is captured in depth can, to a certain degree, be recovered in the RGB images. In contrast, the additional descriptive information which is inherent in RGB, like color information cannot be found in depth images. This will be further analyzed in section 5.5.

The performance obtained for models trained with the two losses is only slightly differing (see Table 1). Cross-modal distillation networks with a baseline trained with softmax loss profit from the deeper neural network architecture Resnet50, while networks with a baseline trained with triplet loss obtain a better result with the shallower Resnet18 architecture. The overall best performance is obtained with a baseline in Resnet50 and softmax loss with an average mAP of 30.1% with RGB as gallery (G) and depth (D) as query and 27.1% for depth as gallery and RGB as query. The corresponding average mAPs for the network with baseline Resnet18 and triplet loss are 28.1% and 27.9%, respectively.

A remarkable finding is the significant difference in performance when alternating the modality used as gallery and query between RGB and depth. Our results suggest that a higher level of performance can be achieved in all networks when the gallery consists of RGB images. The explanation for this behavior can be found in the single-modal re-identification performance of depth and RGB. In fact, when calculating the performance of the network with single-modal re-identification, the RGB modality provides better results than with depth. Therefore, if RGB images are in the gallery the probability of meaningful embeddings for the images is higher than for depth in gallery. As the performance indicators are more influenced by meaningful embeddings in the gallery, we see this effect. Hence, a recommendation for future work on cross-modal re-identification is to report for both gallery and query definitions.

Figure 7 shows an example of results for the best performing cross-modal distillation network (Resnet50 with softmax loss) on BIWI dataset, where the query image is RGB and the gallery image is depth. Query images are selected randomly in test set. This figure highlights the complexity of the task, which is very difficult to solve for humans.

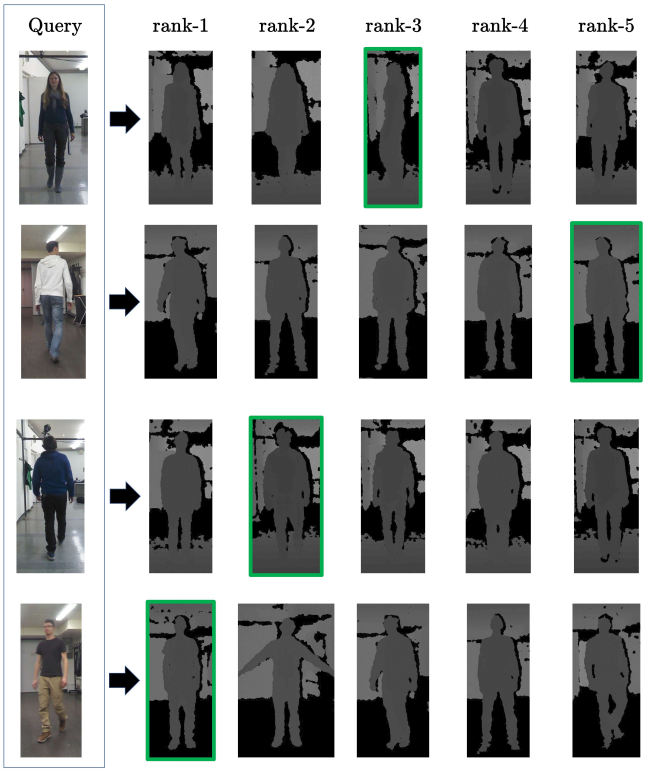


Fig. 7. Example of qualitative results for the proposed architecture on BIWI dataset. The green box denotes the correct match. Gallery (G) and Query (Q) varied for the modalities.

Figure 8 presents the average mAP accuracy of the cross-modal distillation networks trained on the RobotPKU dataset in the cross-modal tasks. We present the same results as with the BIWI dataset. Since it is not possible to train a network with triplet loss in depth (see section 5.1), these results are not reported in the table. The results on RobotPKU data mirror the findings from the BIWI dataset. Again, the transfer from depth to RGB significantly outperforms

the transfer from RGB to depth. The difference of the best networks in mAP is 11%/7.5% for varying query and gallery population. The best overall network is Resnet50 trained with softmax and a transfer from depth to RGB. Similarly to observations on BIWI data, the accuracy for RGB as gallery (G) and depth as query (Q) is higher compared to depth as Gallery (G) and RGB as Query (Q).

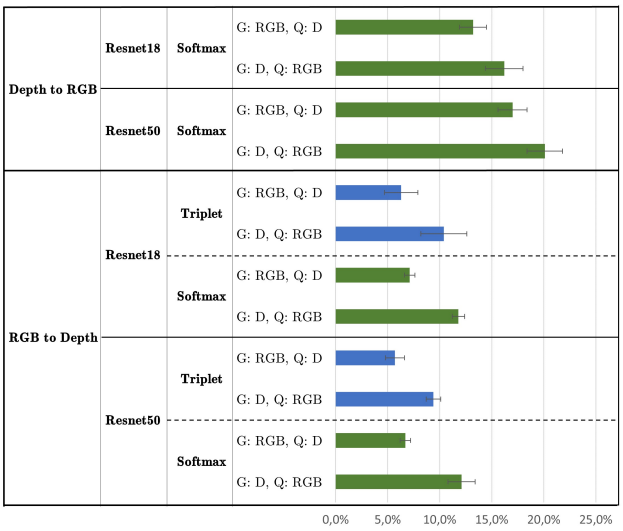


Fig. 8. Average mAP accuracy of various cross-modal distillation networks on the RobotPKU dataset. For all combinations we report varying query (Q) and gallery (G) modalities. The first column indicates the direction of the transfer for the cross-modal distillation.

In summary, to obtain the better results with the cross-modal distillation network, the transfer of knowledge should occur from depth to RGB. As shown in section 5.1 (tables 1 and 2) in the single-modal task a much higher performance was obtained in the RGB modality. Hence, the performance in the single-modal task of the baseline network is not critical to performance for cross-modal distillation. Results suggest that the success of the distillation step is more dependent on the features learned from the modalities. Hence, the features learned in the depth modality were transferable to the RGB modality, while features learned in the RGB modality where not transferable to the depth modality. This gives an indication on the relation between the depth and RGB modality where depth can, to a certain degree, be considered a subset of RGB. The results indicate that networks with a baseline trained with softmax loss and networks with a baseline network in triplet loss obtain similar results. In section 5.3 a more detailed analysis on the influence of the embedding size will be evaluated.

5.3 Sensitivity Analysis for Cross-Modal Distillation

To get a better understanding of the cross-modal distillation network we will present a sensitivity analysis in this chapter. First, the ideal embedding size and layer for the architectures which were identified as best suited for the task in section 5.2 will be analyzed. Second, the influence of the different components of the distillation process will be evaluated.

For the BIWI dataset the best performing cross-modal distillation methods were obtained with a transfer from depth to RGB with a baseline in Resnet50 trained with softmax loss and with a baseline in Resnet18 trained with triplet loss. Hence, for these two methods the influence of differently sized embeddings are analyzed in figure 9 and 10. For the cross-modal distillation network trained with a baseline in softmax loss (figure 9) the two variants of the softmax loss as defined in formulas 3 and 4 are evaluated. The difference between the two definitions is the layer which is defined as the embedding layer. In the variably sized embedding as in formula 3 the features are extracted from the preliminary layer before the softmax loss. For this embedding sizes of 32, 128, 256, 512, 1024 and 2048 are investigated. Embeddings from the novel softmax loss definition as in formula 4 are denoted as classification layer embedding and have a defined size according to the number of training classes, which is 32 for the BIWI dataset and 40 for the RobotPKU dataset. For better visual comparison of the two embedding definitions the obtained performance for the latter are shown as a horizontal line independently of the x-axis. For triplet loss embedding sizes of 32, 128, 256, 512, 1024 and 2048 are evaluated.

It becomes clear that for the BIWI dataset in the single-modal task in pure depth (right graph in figure 9) the networks profit from a bigger size within the preliminary layer embedding up to a convergence. In contrast to that the best performance in the cross-modal tasks after step II. of the cross-modal distillation is obtained when using the classification layer embedding with 37.73% and 39.81% for varying query and gallery definition. For the cross-modal tasks, an optimum for the preliminary layer embedding can be found at a 512 dimensional feature size. However, the results are inferior to the classification layer embedding. The results for a varying embedding size for triplet loss are shown in figure 10. Here, only slight performance variations can be observed for a differing embedding size. The overall best result for the cross-modal task for the BIWI dataset is obtained with the classification layer embedding of the size of training classes, 32. To the best of our knowledge this is the first work identifying the classification layer as a better performing embedding layer than the preliminary layer for a re-identification task. Hence, the suggestion for future work in re-identification is to consider the classification layer embedding as a potential alternative to the preliminary layer embedding.

The results for a varying embedding size for a cross-modal distillation network with a baseline in Resnet50 and softmax loss for the RobotPKU dataset can be seen in figure 11. It gets visible that for all evaluations a clear optimum is reached with an embedding size of 256 with the preliminary layer embedding. In this case, the preliminary layer embedding outperforms the classification layer embedding slightly for all tasks. The best obtained average mAPs for the cross-modal tasks with changing query and gallery are obtained with Resnet50 trained with softmax loss with an preliminary layer embedding of size 256 are 18.13% and 20.52%.

The cross-modal distillation method is highly dependent on a successful knowledge transfer from depth to RGB. To get more insights into this transfer we evaluated the influence on network accuracy in the cross-modal tasks with

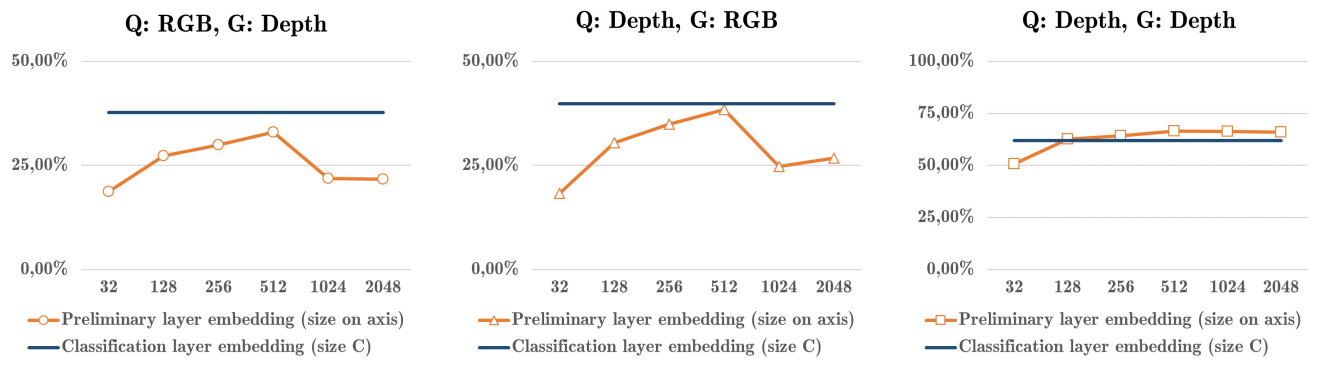


Fig. 9. Analysis of influence of embedding layer and embedding size on the performance of the cross-modal distillation network with Resnet50 and *softmax loss* on the BIWI dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery (left), depth as query and RGB as gallery (middle) and single-modal performance in depth (right).

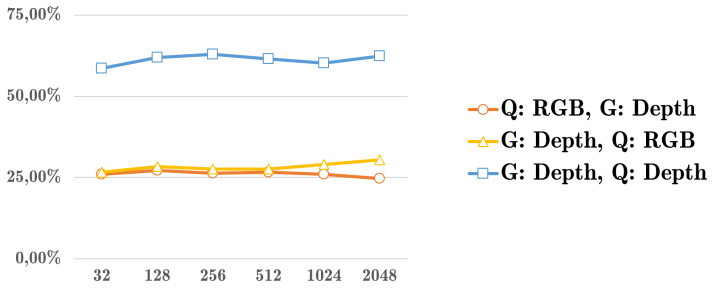


Fig. 10. Analysis of influence of embedding size on the performance of the cross-modal distillation network with Resnet18 and *triplet loss* on the BIWI dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery, depth as query and RGB as gallery, and single-modal performance in depth in the same chart

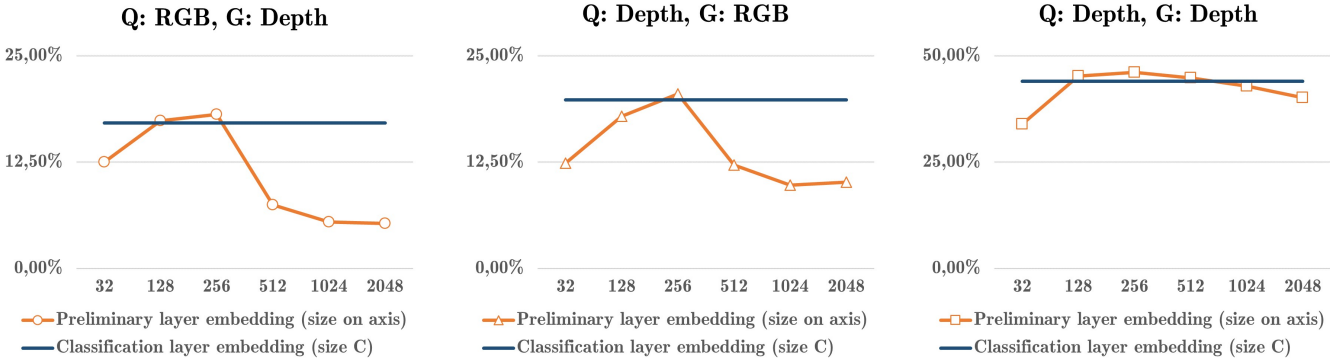


Fig. 11. Analysis of influence of embedding layer and embedding size on the performance of the cross-modal distillation network with Resnet50 and softmax loss on the RobotPKU dataset. Transfer from depth to RGB. Reported are RGB as query and depth as gallery (left), depth as query and RGB as gallery (middle) and single-modal performance in depth (right).

varying components for knowledge transfer. Table 3 shows the impact of copying of weights, and freezing of mid to high-level layers on the accuracy. Results are shown for the BIWI dataset with a cross-modal distillation network with a baseline in Resnet18 trained with classification layer embedding. If the freezing of mid- to high-level layers in the copied network is omitted, performance decreases by 6.8%/3.5%. Another reduction can be seen when the second network is not initialized with the weights of the first network. In this case the cross-modal performance in average mAP decreases by 6.4%/5.8%. These results underline the importance of each component for the cross-modal distillation network in performing knowledge transfer across the modalities.

TABLE 3
Analysis of influence of various training scenarios for knowledge transfer. Results are average accuracy of the BIWI dataset for a cross-modal distillation network using Resnet18 and softmax loss as introduced in formula 4.

Scenario		rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP(%)
No copying of weights, No freezing of layers	Q: RGB, G: D	15.7	49.8	77.9	17.5
	Q: D, G: RGB	19.4	54.6	82.9	23.9
Copying of weights, No freezing of layers	Q:RGB, G: D	22.6	63.1	88.3	23.9
	Q: D, G: RGB	26.9	70.2	91.8	29.7
Copying of weights, Freezing of layers	Q:RGB, G: D	29.8	71.5	91.8	30.6
	Q: D, G: RGB	31.0	73.4	93.1	33.2

TABLE 4
Average accuracy of state-of-the-art and proposed networks for different scenarios on the BIWI dataset. For results from [17] no detailed information on the evaluation procedure was given. As the single-gallery shot is used, this paper reports conservative accuracy indicators a comparison is still possible.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
WHOS, Euclidean [73]	3.2	16.6	31.5	3.7	5.1	18.7	32.6	5.6
WHOS, XQDA [73]	8.4	31.7	50.2	7.9	11.6	34.1	51.4	12.1
LOMO, Euclidean [24]	2.8	16.4	32.5	4.8	3.3	15.6	29.8	5.6
LOMO, XQDA [24]	13.7	43.2	61.7	12.9	16.3	44.8	62.8	15.9
Eigen-depth HOG/SLTP, CCA [17]	8.4	26.3	41.6	-	6.6	27.6	45.0	-
Eigen-depth HOG/SLTP, LSSCDL [17]	9.5	27.1	46.1	-	7.4	29.5	50.3	-
Eigen-depth HOG/SLTP, Corr. Dict. [17]	12.1	28.4	44.5	-	11.3	30.3	48.2	-
Zero-padding network, [10] **Resnet50**	5.86 ± 2.18	25.85 ± 6.35	47.13 ± 8.06	7.28 ± 4.03	10.34 ± 2.68	38.91 ± 6.45	62.84 ± 11.48	9.77 ± 3.80
One-stream network, [10] **Resnet50**	15.68 ± 0.77	50.29 ± 1.18	75.65 ± 0.46	16.86 ± 0.87	19.82 ± 0.33	55.74 ± 0.83	78.92 ± 1.07	23.75 ± 0.30
Cross-modal distillation network, Resnet50, Embedding size 32 (C), (ours)	$\mathbf{34.87 \pm 2.48}$	$\mathbf{75.22 \pm 2.42}$	$\mathbf{93.93 \pm 1.21}$	$\mathbf{35.90 \pm 2.37}$	$\mathbf{36.29 \pm 2.25}$	$\mathbf{77.77 \pm 2.21}$	$\mathbf{94.44 \pm 2.24}$	$\mathbf{38.31 \pm 2.18}$

TABLE 5
Average accuracy of state-of-the-art and proposed architecture for different scenarios on the RobotPKU dataset.

Approach	Query-RGB, Gallery-Depth				Query-Depth, Gallery-RGB			
	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)	mAP (%)
WHOS, Euclidean [73]	3.8	16.3	29.5	3.9	3.5	16.1	31.2	5.4
WHOS, XQDA [73]	10.0	31.8	49.8	8.2	9.8	31.0	48.0	9.8
LOMO, Euclidean [24]	3.6	15.0	28.0	3.9	3.7	15.3	28.7	4.9
LOMO, XQDA [24]	12.9	36.4	56.1	10.1	12.3	37.4	56.1	12.3
Zero-padding network, [10] **Resnet50**	7.76 ± 0.85	29.04 ± 2.57	47.79 ± 3.34	7.67 ± 0.59	6.57 ± 0.64	26.80 ± 2.14	45.62 ± 2.78	8.31 ± 0.56
One-stream network, [10] **Resnet50**	11.92 ± 0.63	38.13 ± 1.01	57.34 ± 2.14	11.42 ± 0.52	12.48 ± 1.01	38.51 ± 1.51	56.77 ± 0.85	14.19 ± 1.37
Cross-modal distillation network, Resnet50, Embedding size 256, (ours)	$\mathbf{19.50 \pm 0.99}$	$\mathbf{50.11 \pm 0.53}$	$\mathbf{67.93 \pm 0.69}$	$\mathbf{18.13 \pm 1.21}$	$\mathbf{21.51 \pm 1.12}$	$\mathbf{54.90 \pm 1.40}$	$\mathbf{72.61 \pm 0.95}$	$\mathbf{20.52 \pm 1.00}$

5.4 Comparison with State-of-the-Art Methods

In this section the results from section 5.2 are taken into a broader scope. Therefore, a comparison to existing methods for cross-modal person re-identification will be taken. Additionally to the presented deep neural network structures for cross-modal person re-identification several methods based on hand-crafted features will be evaluated for the task. Hence, in the following the WHOS feature extractor [73] and the LOMO feature extractor [24] will be investigated. The same features will be extracted for both modalities. The features are compared on basis of Euclidean distance and the additional metric learning step Cross-view Quadratic Discriminant Analysis (XQDA). Additionally, the matching of Eigen-depth and HOG/SLTP features as reported by [17] is included in table 4 for the BIWI dataset.

Table 4 presents the average accuracy of state-of-the-art and proposed networks for different scenarios on the BIWI dataset. First, it is apparent that the hand-crafted feature extractors lead to very low accuracy when matched in the Euclidean space. This is expected, as the modalities depth and RGB are heterogeneous and, hence, no direct comparison of hand-crafted features is possible. When applying the Cross-view Quadratic Discriminant Analysis (XQDA) the performance of the models based on hand-crafted features are significantly enhanced, while the LOMO features lead to the best results. These results also outperform the results from [17] for the Eigen-depth features combined with HOG/SILTP.

Interestingly, also the zero-padding network is outperformed by the conventional approaches. This suggests that the zero-padding with the tested architecture is not suitable for the cross-modal person re-identification task between depth and RGB. For BIWI, the one-stream architecture is outperforming all methods based on hand-crafted features by at least 3%/7% for varying query and gallery in Rank 1 accuracy with a Resnet50 structure. Finally, the cross-modal distillation network enables an additional improvement compared to the one-stream network by 19%/16% for Resnet50.

In table 5 the results for the RobotPKU dataset are shown. Again, the LOMO features with the subsequent metric learning step XQDA obtains the best mAP for the hand-

crafted methods. The one-stream network with Resnet50 structure outperforms LOMO, XQDA in average mAP. The performance increase of the cross-modal distillation network above that of the one-stream network is at 6.7%/6.3%.

Overall results show that the cross-modal distillation network can significantly improve accuracy compared to state-of-the art methods for both BIWI and RobotPKU datasets. This improvement was bigger with BIWI dataset than with the RobotPKU dataset. This is most probably due to the fact, that the BIWI dataset consists of high quality depth images, which are very well synchronized. The depth data in the RobotPKU dataset contains many more flaws like missing limbs and, additionally, the coupled images between depth and RGB are far less synchronized. As the cross-modal distillation network relies on coupled images, poor synchronization of RGB-D images can have have a non-negligible influence on performance. The difficulties in the RobotPKU dataset also explain the lower overall accuracies in RobotPKU in comparison to the BIWI dataset. As all methods based on deep neural networks compared in this section have the same meta-architecture during inference, and were built upon the same feature extractors, the time and memory complexity is the same for all methods. This underlines the superiority of the cross-modal distillation network over the competing methods.

5.5 Analysis of Neural Network Activations

The cross-modal distillation network is state-of-the-art for cross-modal person re-identification. The analysis in section 5.2 showed that the high performance is feasible when transferring knowledge from depth to RGB. To insight into why a baseline trained in the depth modality is that superior, a analysis of deconvolutional images will be made for certain deep neural network architectures. Figure 12 shows deconvolution images for different networks on two images from RGB (a. and c.) and depth (b. and d.) from the BIWI RGBD-ID dataset. The guided backpropagation algorithm was used for visualization of the activations for the networks [40]. The architectures which are shown are separate training for the single-modality task (as in section 5.1), the one-stream network (presented in section 3.2), and our cross-modal distillation method.

The images show that the activations for the different networks are varying considerably. When optimized for the single modalities, the networks in the RGB modality are activated by features inside the torso region of a person, like the color of the same. The network sensing in the depth modality is activated by the outer structure of the torso. For the one-stream network the activation structures are not that clear. For the RGB modality the network is mostly activated by colors of torso and upper legs, while in the depth modality a cluttered outer structure of the torso is captured.

For the RGB modality in the cross-modal distillation network a very different activation map can be observed (images (a) and (c)). Instead of being activated by color features, we see that the network is mostly activated the structure of the torso for those images. Therefore, the knowledge from depth, which is a descriptiveness of the problem with structural details, was transferred to the RGB modality. This finding underlines that the transfer of knowledge between the modalities was successful. As the describing features for the images are similar, the task of embedding to a common feature space is facilitated. This explains the better performance in cross-modal person re-identification as found in section 5.4.

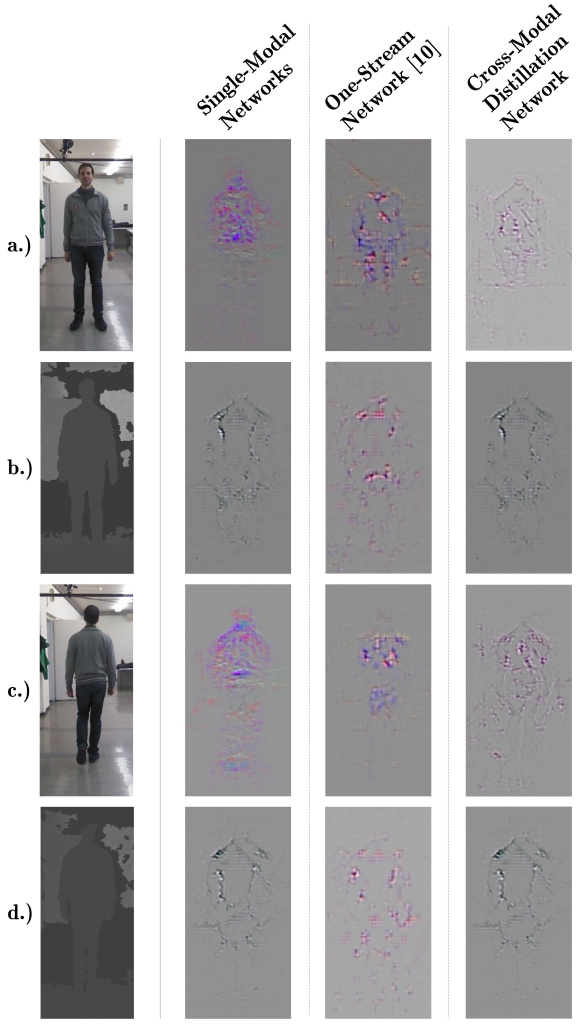


Fig. 12. Comparison of deconvolution images for different networks on BIWI data. Visualization is performed with guided backpropagation [40]. Activation maps of cross-modal distillation network in RGB highly differing to the other techniques.

6 Conclusions

In this work a new technique for cross-modal person re-identification between RGB and depth was presented. The cross-modal distillation network is trained in two steps. Firstly, a deep neural network is optimized in a single-modality with architectures and losses which are proven to be efficient for single-modal person re-identification. In the second step, a distillation of the learned features to the second modality takes place and an embedding of images from both modalities in a common feature space is enforced.

The key difference of our method to the state-of-the-art methods for cross-modal person re-identification with

deep neural networks is its two-step approach. This enables the method to exploit the relation between the two relevant modalities. We find that our transfer-learning approach outperforms state-of-the-art the current state-of-the-art for cross-modal person re-identification between RGB and depth.

Our experiments showed that features which are descriptive in the depth modality can successfully be transferred to the RGB modality for the task of person re-identification. An implication of this is that information captured in depth is to a certain level retrievable in the RGB modality. Following this, we were able to show that for the specific application the depth modality can, up to a certain degree, be considered a subset of the RGB modality. This finding helps to explain the dependence of the RGB modality and the depth modality.

The analysis in this paper also showed that cross-modal person re-identification is a complex task, and the results in absolute numbers suggest that there is still room for improvement. In fact, the accuracies obtained in cross-modal re-identification (tables 4 and 5) are still significantly lower than the accuracies for single-modal re-identification in the more difficult modality (tables 1 and 2). As this is one of the first works concerning the task we want to highlight some potential future directions and current problems in the domain.

First, it will be necessary to obtain bigger datasets to make research more attractive and give data-hungry methods based on deep neural networks the possibility to obtain higher accuracies. The publication of the SYSU-IR dataset in 2017 [10] pushed the interest in cross-modal person re-identification in RGB vs. infrared immensely [11], [12], [13]. A similar effect could be expected for cross-modal re-identification between RGB and depth. Therefore, the amount of persons contained in the datasets would have to rise from less than a hundred for the current datasets to at least the magnitude of several hundreds. Additionally, high-quality depth and RGB images will be necessary. Second, for future research it will be important to expand the considered mode of depth. Especially for the need in intelligent vehicles it will be necessary to evaluate the methods on sparse depth maps, as captured by LiDARs or radars. Therefore, completely new datasets with a high amount of tracked pedestrians and other street objects, will be needed.

Overall, we were able to approach the relevant problem of cross-modal re-identification in RGB and depth for surveillance applications as well as intelligent vehicles in a very effective way. Our method brings the community closer to solve this difficult challenge and our results help to understand the relation between RGB and depth better.

APPENDIX A
SPLIT OF EVALUATION DATASETS

This appendix provides the label of individuals used to form the design (training set plus validation) and test subsets.

A.1 BIWI RGBD-ID dataset:

Design set (Train + Validation set):

0, 1, 4, 5, 6, 7, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20, 25, 26, 34, 35, 38, 39, 40, 43, 50, 56, 57, 58, 59, 61, 62, 65, 66, 67, 69, 70, 73, 74, 76, 77.

Test set:

2, 3, 8, 10, 14, 21, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 36, 37, 41, 42, 44, 45, 46, 47, 48, 49, 51, 52, 53, 54, 55, 60, 63, 64, 68, 71, 72, 75.

A.2 RobotPKU dataset:

Design set (Train + Validation set):

0, 2, 3, 15, 16, 18, 19, 20, 21, 22, 23, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 41, 43, 44, 45, 46, 47, 52, 54, 55, 58, 59, 60, 63, 66, 67, 68, 72, 73, 74, 77, 78, 80, 82, 83, 84, 87, 88.

Test set:

1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 24, 26, 38, 39, 40, 42, 48, 49, 50, 51, 53, 56, 57, 61, 62, 64, 65, 69, 70, 71, 75, 76, 79, 81, 85, 86, 89.

REFERENCES

[1] Gong, S., Cristani, M., Yan, S., Loy and C. C. Eds. "Person re-identification". Springer Science & Business Media, 2014.

[2] Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O. and Radke, R. J., "A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets". In TPAMI, 2018.

[3] Shoubiao T., Feng Z., Li L., Jungong H., Ling S., "Dense Invariant Feature-Based Support Vector Ranking for Cross-Camera Person Reidentification". In IEEE Trans. Circuits Syst. Video Techn., 2018.

[4] Sanping Z., Jinjun W., Jiayun W., Yihong G. and Nanning Z., "Point to Set Similarity Based Deep Feature Learning for Person Re-Identification, In CVPR, 2017.

[5] Xun Y., Meng W., Dacheng T., "Person Re-Identification With Metric Learning Using Privileged Information". In IEEE Trans. Image Processing, 2018.

[6] Wang, K., Yin, Q., Wang, W., Wu, S. and Wang, L. "A comprehensive survey on cross-modal retrieval". In arXiv preprint arXiv:1607.06215, 2016.

[7] Wang, M. and Weihong D.. "Deep Visual Domain Adaptation: A Survey". In Neurocomputing, 2018.

[8] Gupta, S., Judy H. and Jitendra M. "Cross modal distillation for supervision transfer". In CVPR, 2016.

[9] Zheng, L., Yi Y. and Alexander G. H.. "Person re-identification: Past, present and future". arXiv:1610.02984 2016.

[10] Wu, A., Zheng, W. S., Yu, H. X., Gong, S. and Lai, J. "RGB-infrared cross-modality person re-identification". In ICCV, 2017.

[11] Ye, M., Lan, X., Li, J. and Yuen, P. C. "Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification". In AAAI, 2018.

[12] Ye, M., Wang, Z., Lan, X. and Yuen, P. C "Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking". In IJCAI, 2018.

[13] Dai, P., Ji, R., Wang, H., Wu, Q. and Huang, Y. "Cross-Modality Person Re-Identification with Generative Adversarial Training". In IJCAI, 2018.

[14] Lohani, B., Chacko, S., Ghosh, S., Sasidharan and S. "Surveillance system based on Flash LiDAR". In International Congress on Cartography for Sustainable Earth Resource Management. Vol. 32. 2013.

[15] Sudhakar, P., Anitha Sheela, K. and Satyanarayana, M. "Imaging Lidar system for night vision and surveillance applications". In ICACCS, 2017.

[16] Wu, A., Wei-Shi Z. and Jian-Huang L.. "Robust depth-based person re-identification". Transactions on Image Processing, 2017.

[17] Zhuo, J., Zhu, J., Lai, J. and Xie, X. "Person Re-identification on Heterogeneous Camera Network". CCF Chinese Conference on Computer Vision, 2017.

[18] Munaro, M., Fossati, A., Basso, A., Menegatti, E. and Van Gool, L. "One-shot person re-identification with a consumer depth camera". Person Re-Identification, 2014.

[19] Liu, H., Liang H. and Liqian M.. "Online RGB-D person re-identification based on metric model update". CAAI Transactions on Intelligence Technology 2.1 (2017): 48-55

[20] Farenzena, M., Bazzani, L., Perina, A. Murino, V. and Cristani, M. "Person re-identification by symmetry-driven accumulation of local features". In CVPR, 2010.

[21] Bhuiyan, A., Perina, A., Murino, V. "Person re-identification by discriminatively selecting parts and features". In ECCVWK, 2014.

[22] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, "Custom pictorial structures for re-identification". In BMVC, 2011.

[23] Panda, R., Bhuiyan, A, Murino, V. and Roy-Chowdhury, A. K. "Unsupervised Adaptive Re-identification in Open World Dynamic Camera Networks". In CVPR, 2017.

[24] Liao, S., Hu, Y., Zhu, X. and Li, S. Z. "Person re-identification by local maximal occurrence representation and metric learning". In CVPR, 2015.

[25] Wu, Z., Li, Y., Radke, Richard J. "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features". In TPAMI, 2015.

[26] Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., Smith, J. R. "Learning locally-adaptive decision functions for person verification". In CVPR, 2013.

[27] Liu, C., Gong, S., Loy, C. C. "On-the-fly feature importance mining for person re-identification". In Pattern Recognition, 2014.

[28] Liao, S., Li, Stan Z. "Efficient psd constrained asymmetric metric learning for person re-identification". In ICCV, 2015.

[29] Liao, S., Li, Stan Z. "Learning to rank in person re-identification with metric ensembles". In CVPR, 2015.

[30] Liao, S., Hu, Y., Zhu, X., Li, Stan Z. "Person re-identification using kernel-based metric learning methods". In ECCV, 2014.

[31] Pan, S. J., Yang, Q. "A survey on transfer learning". In IEEE Transactions on knowledge and data engineering, 2010.

[32] Weiss, K., Khoshgoftaar, T. M, Wang, D.. "A survey of transfer learning". In Journal of Big Data, 2016.

[33] Duan, L., Xu, D., Tsang, I. "Learning with augmented features for heterogeneous domain adaptation". In arXiv preprint arXiv:1206.4660, 2012.

[34] Zhu, Y., Chen, Y., Lu, Z., Pan, S. J., Xue, G., Yu, Y., Yang, Q. "Heterogeneous Transfer Learning for Image Classification". In AAAI, 2011.

[35] Harel M, Mannor S. "Learning from multiple outlooks". In ICML, 2011.

[36] Ma, Z., Yang, Y., Nie, F.,Sebe, N., Yan, S., Hauptmann, A. G. "Harnessing lab knowledge for real-world action recognition". In IJCV, 2014.

[37] Yang, Y., Ma, Z., Xu, Z., Yan, S., Hauptmann, A. G. "How related exemplars help complex event detection in web videos?". In ICCV, 2013.

[38] Jie, L., Tommasi, T., Caputo, B. "Multiclass transfer learning from unconstrained priors". In ICCV, 2011.

[39] Gopalan, R., Li, R., Chellappa, R. "Domain adaptation for object recognition: An unsupervised approach".. In ICCV, 2011.

[40] Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M. "Striving for simplicity: The all convolutional net". In ICLR (workshop track), 2015.

[41] He, K., Zhang, X., Ren, S., Sun, J. "Deep residual learning for image recognition". In CVPR, 2016.

[42] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. "Salient color names for person re-identification". In ECCV, 2014.

[43] N. Gheissari, T. B. Sebastian, and R. Hartley. "Person reidentification using spatiotemporal appearance". In CVPR, 2006.

[44] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. "Boosted human re-identification using riemannian manifolds". In Image and Vision Computing, 2012.

[45] D. Gray and H. Tao. "Viewpoint invariant pedestrian recognition with an ensemble of localized features". In ECCV, 2008.

[46] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. "Person re-identification by support vector ranking". In BMVC, 2010.
[47] R. Schwartz and L. S. Davis. "Learning discriminative appearance-based models using partial least squares". In Symposium on Computer Graphics and Image Processing, 2009.
[48] E. Corvee, F. Bremond, M. Thonnat, et al. "Person re-identification using spatial covariance regions of human body parts". In AVSS, 2010.
[49] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. "Person re-identification in multicamera system by signature based on interest point descriptors collected on short video sequences". International Conference on Distributed Smart Cameras, 2008.
[50] B. Ma, Y. Su, and F. Jurie. "Bicov: a novel image representation for person re-identification and face verification". In BMVC, 2012.
[51] M. Guillaumin, J. Verbeek, and C. Schmid. "Is that you? Metric learning approaches for face identification". In ICCV, 2009.
[52] K. Q. Weinberger, J. Blitzer, and L. K. Saul. "Distance metric learning for large margin nearest neighbor classification". In NIPS, 2006.
[53] Koestinger, M., Martin H., Paul W., Peter M. R., and Horst B. "Large scale metric learning from equivalence constraints". In CVPR, 2012.
[54] Giuseppe L. , Iacopo M. , Alberto D. B., "Matching People across Camera Views using Kernel Canonical Correlation Analysis, International Conference on Distributed Smart Cameras, 2014.
[55] Yi, D., Lei, Z., Liao, S., Li, S.Z., "Deep Metric Learning for Person Re-identification". In CVPR, 2014.
[56] Ahmed, E.; Jones, M.; Marks, T.K. "An improved deep learning architecture for person re-identification". In CVPR, 2015.
[57] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N. "Person re-identification by multi-channel parts-based CNN with improved triplet loss function". In CVPR, 2016.
[58] Varior, R.R., Haloi, M., Wang, G. "Gated siamese convolutional neural network architecture for human re-identification". In ECCV, 2016.
[59] Xiao, T., Li, H., Ouyang, W., Wang, X. "Learning deep feature representations with domain guided dropout for person re-identification". arXiv, 2016.
[60] Hermans, A., Lucas B., and Bastian L. "In defense of the triplet loss for person re-identification". arXiv preprint arXiv:1703.07737, 2017.
[61] Chen, W., Xiaotang C., Jianguo Z., and Kaiqi Huang. "Beyond triplet loss: a deep quadruplet network for person re-identification". In CVPR, 2017.
[62] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. "Embedding deep metric for person re-identification: A study against large variations". In ECCV, 2016.
[63] Ristani, E., and Carlo T. "Features for Multi-Target Multi-Camera Tracking and Re-Identification". arXiv preprint arXiv:1803.10859 (2018).
[64] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang. "Joint learning of single-image and cross-image representations for person re-identification". In CVPR, 2016.
[65] D. Li, X. Chen, Z. Zhang, K. Huang. "Learning deep context-aware features over body and latent parts for person re-identification". In CVPR, 2017.
[66] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. "Going deeper with convolutions". In CVPR, 2015.
[67] Zheng, Liang, et al. "Person Re-identification in the Wild". In CVPR, 2017.
[68] Zheng, Z., Liang Z., and Yi Y. "A discriminatively learned cnn embedding for person reidentification". In transactions on Multimedia Computing, Communications, and Applications, 2017.
[69] L. Yutian, et al. "Improving person re-identification by attribute and identity learning". arXiv preprint arXiv:1703.07220, 2017.
[70] M. Geng, Y. Wang, T. Xiang, and Y. Tian. "Deep transfer learning for person re-identification". In arXiv:1611.05244, 2016.
[71] Li, Y. J., Yang, F. E., Liu, Y. C., Yeh, Y. Y., Du, X., Wang, Y. C. F. "Adaptation and Re-Identification Network: An Unsupervised Deep Transfer Learning Approach to Person Re-Identification". arXiv:1804.09347, 2018.
[72] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. "Semisupervised multi-feature learning for person re-identification". In AVSS, 2013.
[73] Lisanti, G., Masi, I., Bagdanov, A. D., Del Bimbo, A.. "Person re-identification by iterative re-weighted sparse ranking". In TPAMI, 2015.
[74] Pala, F., Satta, R., Fumera, G., Roli, F., "Multimodal person re-identification using RGB-D cameras". In Transactions on Circuits and Systems for Video Technology, 2016.
[75] Mogelmose, A., Bahnsen, C., Moeslund, T. B., Clapes, A. and Escalera, S. "Tri-modal person re-identification with rgb, depth and thermal features". In CVPR, 2013.
[76] John, V., Englebienne, G. and Krose, B., "Person re-identification using height-based gait in color depth camera". In ICIP, 2013.
[77] R. Satta, F. Pala, G. Fumera, and F. Roli, "Real-time appearance-based person re-identification over multiple kinect cameras. In VISAPP, 2013.
[78] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q. "Scalable person re-identification: A benchmark". In ICCV, 2015

Frank Hafner obtained a Bachelor degree in Industrial Engineering and Management from Karlsruhe Institute of Technology (KIT), Germany, in 2016. He received his Master's degree from Technical University of Delft, Netherlands, in Vehicle Engineering with the specialization in Perception and Modelling in 2018. For his thesis project he collaborated with Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), Montréal, Canada. He is currently working as a research and development engineer for autonomous driving at ZF Friedrichshafen AG. His main research interests include re-identification of objects in different contexts, RGB-D vision and efficient deployment of neural networks.

Amran Bhuiyan received the Bachelor degree in Applied Physics, Electronic & Communication Engineering from University of Dhaka, Bangladesh in 2009, the M.Sc. degree in Computer Engineering and Information Technology from the Lucian Blaga University of Sibiu, Romania under the Erasmus Mundus external window in 2011 and the Ph.D. degree in Pattern Analysis and Computer Vision from the Istituto Italiano di Tecnologia, Genova, Italy. He is currently a Postdoctoral Researcher with LIVIA, École de Technologie Supérieure, Université du Québec, Montréal, Canada. His main research interests include computer vision, machine learning, person re-identification and video surveillance.

Julian F.P. Kooij (M'08) obtained the PhD degree in artificial intelligence at the University of Amsterdam in 2015, where he worked on unsupervised and predictive models of pedestrian behaviour. In 2013 he interned at Daimler AG in Germany to apply his research to path prediction of vulnerable road users for highly-automated vehicles. From 2014-2016 he was as a PostDoc at the computer vision lab of the EEMCS faculty of the TU Delft, developing RGB-D vision techniques to detect body motions of patients with neurological disorders, and collaborated on the Technology In Motion lab at Leiden University Medical Hospital. Since 2016, he is an Assistant Professor at the Intelligent Vehicles group, part of the Cognitive Robotics department at the TU Delft. His current research interests include semi-supervised machine learning, Bayesian inference, and predictive motion models, applied to environment perception for automated driving in cluttered urban environments.

Eric Granger (M'00) received the Ph.D. degree in Electrical Engineering from the Ecole Polytechnique de Montreal, Montreal, QC, Canada, in 2001. He was a Defense Scientist with Defence Research and Development Canada-Ottawa, Ottawa, ON, Canada from 1999 to 2001. He was with Mitel Networks, Ottawa, from 2001 to 2004, where was involved in research and development. In 2004, he joined the École de technologie supérieure (Université du Québec), Montreal, where he is currently a Full Professor and the Director of Laboratoire d'imagerie, de vision et d'intelligence artificielle, Montreal, a research laboratory focused on computer vision and artificial intelligence. His current research interests include adaptive pattern recognition, machine learning, computer vision, and computational intelligence, with applications in biometrics, face recognition and analysis, video surveillance, and computer/network security.