Person Re-identification Using Spatial and Layer-Wise Attention

Lejbølle, Aske Rasch; Nasrollahi, Kamal; Krogh, Benjamin; Moeslund, Thomas B.

Published in:
I E E E Transactions on Information Forensics and Security

DOI (link to publication from Publisher):
10.1109/TIFS.2019.2938870

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Person Re-identification Using Spatial and Layer-Wise Attention

Aske R Lejbølle, Kamal Nasrollahi, Benjamin Krogh and Thomas B Moeslund

***Abstract*—Person re-identification requires extraction of discriminative features to ensure a correct match; this must be done independent of challenges, such as occlusion, view, or lighting changes. While occlusion can be eliminated by changing the camera setup from a horizontal to a vertical (overhead) viewpoint, other challenges arise as the total visible surface area of persons is decreased. As a result, methods that focus on the most discriminative regions of persons must be applied, while different domains should also be considered to extract different semantics. To further increase feature discriminability, complementary features extracted at different abstraction levels should be fused. To emphasize features at certain abstraction levels depending on the input, fusion should be done intelligently. This work considers multiple domains and feature discrimination, where a multimodal convolution neural network is applied to fuse RGB and depth information. To extract multi-local discriminative features, two different attention modules are proposed: (1) a spatial attention module, which is able to capture local information at different abstraction levels, and (2) a layer-wise attention module, which works as a dynamic weighting scheme to assign weights and fuse local abstraction-level features intelligently, depending on the input image. By fusing local and global features in a multimodal context, we show state-of-the-art accuracies on two publicly available datasets, DPI-T and TVPR, while increasing the state-of-the-art accuracy on a third dataset, OPR. Finally, through both visual and quantitative analysis, we show the ability of the proposed system to leverage multiple frames, by adapting feature weighting depending on the input.**

***Index Terms*—Artificial neural networks, dynamic feature fusion, multimodal sensors, person re-identification, soft attention**

I. Introduction

SINCE the beginning of the new millennium, person re-identification (re-id) has seen increased interest in the research community as the topic is perceived as both difficult and important [1]–[4]. Identification and verification involve matching an unknown signature to a database of either a single known or multiple known signatures. Re-id is the task of matching an anonymous signature to a database of anonymous signatures to find a correct match. Within computer vision, this is accomplished by matching signatures, i.e., features of a person extracted from images or a video in one camera view to features of persons extracted from images or a video in another. Features can, for example, contain hand-engineered low-level color and texture information [4]–[8], which can be extracted from small body patches, body parts, or the entire body, or they can contain high-level information by encoding low-level features using sparse coding [9]–[11]. Features are matched using a predefined metric, such as Euclidean distance, although, to increase the accuracy of the system, supervised metric learning [5], [12]–[15] is often considered to maximize the distances between non-matching feature pairs and minimize the distances between matching ones. In recent years, however, Convolutional Neural Networks (CNNs) [16]–[21] have become increasingly popular due to their ability to learn discriminative high-level features by combining feature learning and classification in an end-to-end training scheme.

Due to increased data requirements when training CNNs, larger re-id datasets have been published in the last few years [22]–[24]. These datasets are more realistic in terms of the number of deployed cameras and environmental changes between views. A common characteristic of these datasets is the viewpoint, which is primarily horizontal and allows occlusions and changing views, as shown in Figure 1 (a). To eliminate these challenges, the position of the camera can be changed to a vertical (overhead) viewpoint, as shown in Figure 1 (b). In this work, only data captured from an overhead viewpoint is considered. Changing the viewpoint does, however, also increase the probability of removing important textural information from either the clothing or the face of a person. To counter the decrease in visual information, feature discriminability can be increased by adding additional information from other modalities. In connection with the overhead viewpoint, we add depth information when devising novel features. Adding depth enables us to model the height or body part ratios of persons, which can be used to learn a multimodal feature representation based on both RGB and depth modalities.

Additionally, the use of local feature representations has shown to outperform global ones [4]. In case of hand-engineered features, this is done by sampling small-image patches, typically of size 10×10 pixels, and extracting features from each patch. In terms of a CNN, this can be achieved by learning part-specific networks, either by splitting the body horizontally into a predefined number of body parts [18], [20] or by using part localization algorithms [21], [26]. Another option is to apply a soft attention mechanism [27]

Manuscript received February 13, 2019; revised March 30, 2019; revised July 26, 2019; accepted August 28. This work is supported by Innovation Fund Denmark under Grant 5189-00222B. *(Corresponding author: Aske Rasch Lejbølle.)*

A. R. Lejbølle is with the Visual Analysis of People laboratory in the Department of Architecture, Design and Media Technology at Aalborg University, Rendsburggade 14, 9000 Aalborg, and with Veovo, Hækken 2, 9310 Vodskov, Denmark (e-mail: asrl@create.aau.dk/aske.lejboelle@veovo.com).

B. Krogh is with Veovo, Hækken 2, 9310 Vodskov, Denmark (e-mail: benjamin.krogh@veovo.com).

K. Nasrollahi and T. B. Moeslund are with the Visual Analysis of People laboratory in the Department of Architecture, Design and Media Technology at Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark (e-mail: {kn,tbm}@create.aau.dk).

(a) (b)

Fig. 1. (a) Example of a horizontal viewpoint with changing views causing differences in appearance and partly occlusion [22]. (b) Example of an overhead viewpoint eliminating issues in (a) [25].

with the purpose of extracting features from only a single, or few, discriminative regions in the input image. Different from horizontally splitting the body or applying localization algorithms, the attention mechanism captures information from local regions based on a learning scheme. This is based on calculating a two-dimensional weight matrix, which works as a mask on the input, where each element represents a weight in the interval [0,1] and is learned through back propagation. This will be referred to as *spatial attention*, which is applied in [28] to determine the importance of spatial locations at different layers of a neural network based on fusion of RGB and depth features. Since different layers of a CNN produce features at different abstraction levels [29], features produced by spatial attention modules represent local context information at different abstraction levels. To take advantage of complementary low-, mid-, and high-level information, features at different abstraction levels are often fused, simply by concatenation. Concatenation of features causes all elements in the resulting feature vector to be weighted equally, which is inexpedient if features at one abstraction level contain noisy information. Furthermore, local features from certain layers might be unnecessary and finding the optimal combination of relevant local features is impractical and time consuming if a very deep neural network is implemented. Instead, features should be fused using a dynamic weighting scheme that considers relevance to properly weight local features in order to maximize accuracy.

In this work, we introduce a multimodal dynamic weighting scheme as a layer-wise attention module to weight the output features of several spatial attention modules, based on the input. Since we focus on images captured from an overhead viewpoint which, depending on the height of the camera position, results in a more narrow view, sufficient data of each person might not be captured to properly exploit video-based methods, such as Recurrent Neural Networks (RNN). As a result, in this work, we consider only image-based models to learn a multimodal representation. Given a CNN consisting of L convolution layers, each convolution layer, in practice, can be followed by a spatial attention module and thus produce features that contain local context information. Instead of simply concatenating the features, the layer-wise attention module dynamically apply weights, which are learned by a learning scheme, to each feature vector and summarizes the outputs to a single discriminative feature vector. Thus, we end up with a multi-local context feature vector, which is a weighted summary of local context features at different abstraction levels. To take advantage of complimentary local and global information, multi-local context features from each modality are fused with high-level global features; these are referred to as multi-level features. Finally, multi-level modality features from RGB and depth, respectively, are fused to a multi-level multimodal feature representation. In summary, the main contributions include:

- A layer-wise attention module used to dynamically assign weights to local context features at different abstraction levels, depending on the input.
- An analysis of the output of the spatial and layer-wise attention modules used to reason how the data affects the weighting of features at different abstraction levels.
- A demonstration that a combination of spatial and layer-wise attention in a multimodal context provides state-of-the-art accuracy on several datasets collected from an overhead viewpoint.

The rest of the paper is structured as follows. Related work is outlined in Section II, which is followed by a description of the proposed system in Section III, including the baseline architecture as well as spatial and layer-wise attention modules. In Section IV, experimental results are presented along with ablation studies and an analysis of the proposed attention modules. In addition, a comparison between the proposed system and state-of-the-art systems is presented. Finally, the work is concluded in Section V.

II. Related Work

A. CNN in Person Re-Identification

Since the development of early CNNs for the purpose of re-id, part-based learning has been considered with the aim to capture more discriminative local features. Yi et al. [20] proposed a CNN consisting of three separate streams, each processing an image that is split into a similar number of overlapping parts. Part-based features are then fused in a fully connected layer before classification. A similar approach was proposed by Cheng et al. [17], who split the body into four parts to learn part-specific features that are fused with full-body features. Similar to [20], features are fused late in the network. A more sophisticated part-based model was proposed by Ustinova et al. [18], who trained three part-specific sub-networks; instead of using a single sub-network per body part, they trained multiple sub-networks, and fused part-based features from corresponding sub-networks by a bilinear operation to retain geometric information in the input image.

More recently, Zhao et al. [21] proposed an architecture that consists of a Region Proposal Network (RPN) to locate 14-body joints in order to extract seven sub-regions of the body. Part-specific sub-networks were trained based on each of the sub-regions, followed by a Feature Fusion Network (FFN). This approach fuses part-specific information in a pyramid structure. Part-specific learning by joint localization was also proposed by Li et al. [30]; rather than applying an RPN, they

applied a Spatial Transformer Network (STN) to locate head-shoulder, upper-body, and lower-body regions.

Instead of training part-specific sub-networks, Suh et al. [31] trained a part map extractor to capture features from different body parts. Combined with an appearance map extractor to capture appearance features, a bilinear pooling operator is applied to fuse the two feature types, resulting in a part-aligned feature representation. Finally, Sarfraz et al. [26] proposed an architecture that takes as input a 17-channel image, including RGB and 14 keypoint channels containing keypoint locations. Furthermore, a separate view predictor was trained to model view information by calculating weights as probabilities of being 'front', 'back', or 'side'.

The proposed part based models [17], [18], [20] are able to capture and fuse local information from different body parts, and by adding a view predictor [26], it is even possible to add invariance to rotational changes. However, this assumes that images are captured from a horizontal viewpoint. If an overhead viewpoint is considered, certain body parts will be less visible, which will make it more difficult to achieve a proper result.

B. RGB-D CNN Models

Multimodal RGB-D CNNs have been proposed to a variety of applications [32]–[35]. For object recognition, Eitel et al. [32] proposed a two-stream CNN to fuse high-level RGB and depth features by adding a fully connected layer late in the network. An RGB-D CNN was proposed in [33] for pose estimation by, in a similar manner, processing RGB and depth images individually and using fused RGB-D features to train an SVM to determine the pose of objects. McCormac et al. [35] proposed a semantic mapping network, where depth is added as a fourth channel in the input to train an RGB-D semantic segmentation network.

Within re-id, the majority of published work focus on either RGB or depth, while fusion of the two modalities is rarely considered. Hand-crafted RGB-D features were devised by Liciotti et al. [25], who fused low-level color features from HSV histograms with anthropometric features extracted in the depth domain. Hand-crafted features were also devised by Wu et al. [36], who fused rotation invariant Eigen-depth features with low-level patch-based color and texture features. In the case of deep neural models, Karianakis et al. [37] proposed a combined CNN and LSTM to learn spatiotemporal depth features based on low-level knowledge transfer between an RGB and depth CNN. Additionally, they exploited frame-level weights by adding a Reinforced Temporal Attention (RTA) module, which infers the importance of each frame in a sequence using a hard attention mechanism, which has previously been introduced for image captioning [38]. Additionally, they considered fusion of spatiotemporal depth features and RGB features extracted from a CNN that was trained on upper body images of persons. To our knowledge, the only other model within re-id to consider RGB-D features from a CNN is the work of [39], in which RGB and depth images are processed by modality- based sub-networks, while corresponding features are fused by concatenation in fully connected layers late in the network.

While [39] does not consider attention to capture local context features, [37] applies a coarse frame-level attention mechanism that does not capture and weight local information. Our proposed system does not only consider fusion of global RGB and depth features, but it also adds an attention mechanism to dynamically fuse local context features to consider complementary multi-level features.

C. Attention in Person Re-identification

An increasing number of CNNs that apply attention are being proposed in the field of re-id. Inspired by attention, Zhao et al. [40] proposed an architecture that uses part map detectors to estimate two-dimensional weight matrices that are multiplied by an input to output part feature maps. Here, the part map detectors are implemented as 1×1 convolutions followed by a sigmoid activation. A Comparative Attention Network (CAN) was proposed by Liu et al. [41], in which attention is applied to dynamically capture 'glimpses', i.e., minor regions in the input, by calculating spatial weight matrices used as masks on the input image. The dynamic element is added using a Long-Short Term Memory (LSTM) layer, which considers the masked input using the mask at the previous time step, along with the previous hidden state, and outputs a weight matrix, which attends a different area in the input. Masks were also generated by Song et al. [42], who proposed a mask-guided contrastive attention model consisting of three streams; one that learns from the regular input, and two others that learn from a foreground body and background image. The image is segmented by considering an additional binary mask in the input; this is combined with an attention loss to guide the attention map generation used for segmentation. Si et al. [43] proposed a Dual ATtention Matching network (DuATM), which learns a dual attention mechanism to match aligned feature pairs from an input triplet. Distances are then aggregated by average pooling and used with a triplet loss to update the network weights.

Li et al. [44] considered the misalignment issue in the input by proposing a Harmonious Attention CNN (HA-CNN), which combines regional attention [45], spatial attention [46], and channel attention [47] to capture both fine-grained pixel information at global level and coarse discriminative regions at local level. In case of depth modality, Haque et al. [48] proposed a recurrent attention model (RAM), which combines a localization network to capture glimpses with a CNN+LSTM to extract spatiotemporal features from the regions. Attention has been applied also in the context of video-based re-id [49]. This will not be described in detail in the present work as the focus is only on image-based models.

Despite being able to learn fine-grained masks that are applied to the input, [41] adds additional complexity to the model by implementing attention using an LSTM, while [42] requires additional binary ground truth masks during training. Meanwhile, [44] fuses features from different local regions simply by concatenation before propagating the fused feature to a fully connected layer, hence, it does not consider the importance of each local region. Finally, previous work consider either RGB or depth as input during model training.

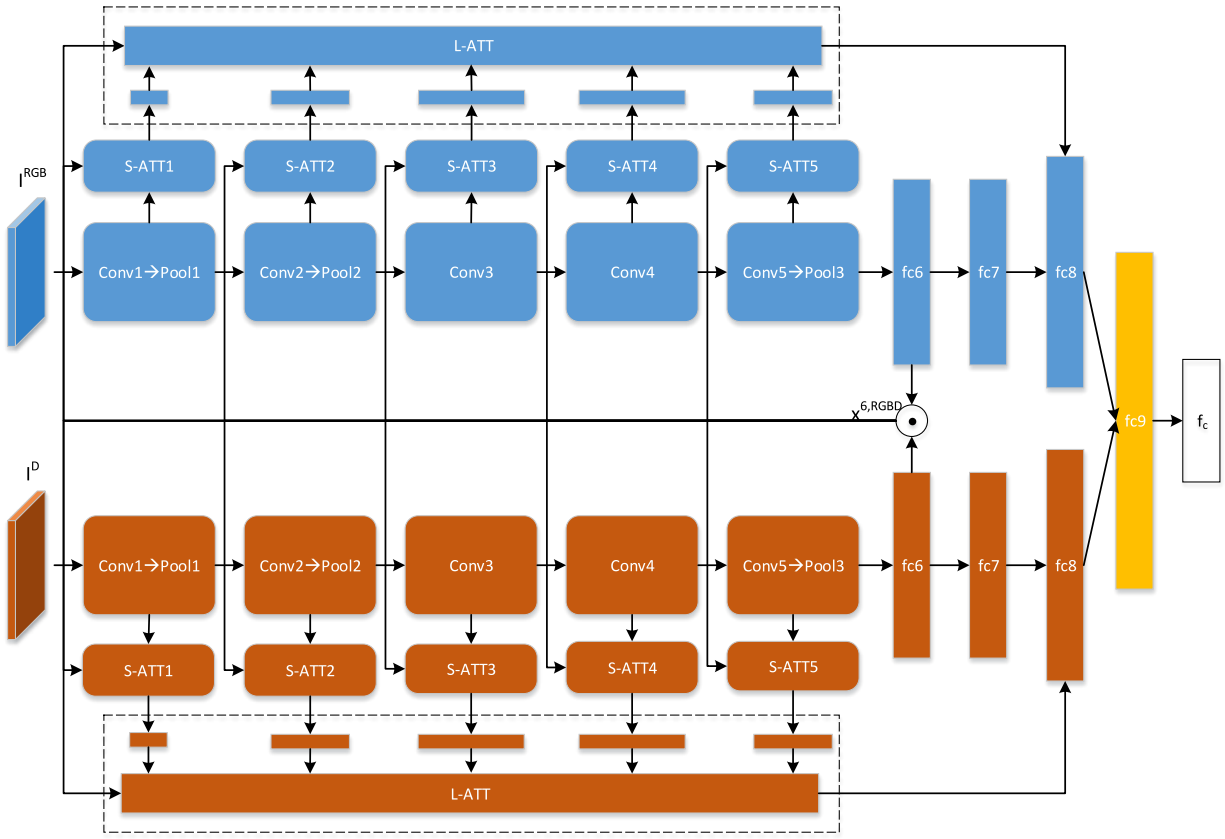

Fig. 2. Overview of the Spatial and Layer-wise ATTention network (SLATT). RGB and depth images, respectively, are fed to separate CNNs pre-trained on modality dependent data. Each convolution layer in the network forward propagates the output, both to the next layer in the network and to S-ATT modules that calculate spatial attention features using RGB- and depth-based features, $x^{6,RGBD}$. Here, $x^{6,RGBD}$ is the resulting feature from fusing the rgb and depth features of the sixth network layer (fc6), respectively, by multiplication. This is indicated by the $\odot$ symbol. Outputs of S-ATT modules are propagated to the L-ATT module, which calculates attention-based feature weights. Modality-based local and global features are fused in $fc8$, while multi-level RGB and depth are fused in $fc9$. Finally, classification is performed in f_c.

To our knowledge, the only previous work to consider multimodal attention is the Multimodal ATtention network (MAT) [28]. In this work, spatial attention weights are calculated for different layers of a CNN based on fusion of features from different modalities. Extracted local features are fused with global ones, and, finally, RGB and depth features are fused to a multimodal feature in the last layer of the network.

D. Dynamic Feature Fusion

Dynamic feature fusion has been studied mostly in connection with the fusion of multiple modalities [50]–[52]. To describe videos, Zhang et al. [50] proposed a combination of appearance and motion features from video clips that are dynamically fused by a weighted summary, where weights are calculated by applying an attention mechanism. The attention mechanism takes as input the motion or appearance feature, along with the intermediate hidden state from a decoder LSTM, to model the relevance of the feature. In video classification, Long et al. [52] proposed an Attention Fusion scheme in which RGB, flow, and audio features are fused by applying a Bidirectional LSTM, which models dependencies between modalities and, based on this, output global modality-based representations that are fused by concatenation. Attention-based dynamic feature fusion was also proposed for video description by Hori et al. [51], who applied a soft attention scheme using the previous hidden state of an LSTM decoder along with a modality feature to output a multimodal attention weight. Our proposed dynamic weighting scheme is mostly similar to the work of [51]. However, rather than dynamically fusing features from different modalities, we leverage multimodal information to fuse local abstraction level features for each modality.

Dynamic fusion of features at different abstraction levels has not often been considered [53], [54]. In case of human pose estimation, Chu et al. [53] proposed an 8-stack hourglass network, where each stack outputs multi-resolution attention maps that are fused by summation and applied to the output of the stack. Furthermore, Chen et al. [54] proposed an RGB-D object detection network by introducing an Attention-ware Cross-modal Cross-level Fusion (ACCF) module, which concatenates RGB and depth feature maps and calculates channel-wise weights to model dependencies between RGB and depth channels. By propagating the output of an ACCF module late in the network back to lower layers of the network, predictions are generated in a coarse-to-fine manner.

In re-id, feature fusion most often is done by concatena-

tion, as described in Section II-B. In [44], local information generated from soft attention mechanisms is fused with global information generated from a hard attention mechanism; this is done by tensor addition to increase interaction. Lastly, Chang et al. [55] proposed the Multi-Level Factorization Net (MLFN), which consists of multiple blocks at different abstraction levels, each calculating a weighted summary of outputs from sub-networks in the given block. Weights from all blocks are additionally fused with features from the last block by an average operation. The former model only leverages RGB information to calculate attention weights, whereas the latter models use multi-level semantics by averaging multi-level features with a high-level feature representation. Here, we consider multimodal features to dynamically model and fuse features at different abstraction levels.

III. Proposed System

The proposed system is shown in Figure 2. Given a pair of RGB and depth images, the system extracts multi-local context features by dynamically assigning weights to local context features at different abstraction levels. This is achieved by implementation of two attention modules: one that models the importance of spatial locations within feature maps at different abstraction levels (S-ATT), and another that models the importance of abstraction-level features (L-ATT). The output of the L-ATT is a feature vector containing local discriminative information, which is fused with a feature vector containing global information; this results in multi-level RGB and depth features. The two modality-based multi-level features are fused to generate a multimodal feature vector that is used for re-identification. The entire system is summarized in Table I; superscripts are neglected for simplicity. In the remaining part of the paper, we will refer to this system as *SLATT*.

TABLE I
Overview of the SLATT architecture, including output sizes from the S-ATT modules at each abstraction level. M denotes the number of persons in the training set. Similar structures are used to process RGB and depth images, respectively, and output multi-level features of similar sizes from fc8. This is indicated by (***x2***).

Layer	*Output size*	*S-ATT output size*
Input	227×227×3 (x2)	
Conv1	55×55×96 (x2)	
Pool1	27×27×96 (x2)	1×96 (x2)
Conv2	27×27×256 (x2)	
Pool2	13×13×256 (x2)	1×256 (x2)
Conv3	13×13×384 (x2)	1×384 (x2)
Conv4	13×13×384 (x2)	1×384 (x2)
Conv5	13×13×256 (x2)	1×256 (x2)
Pool3	6×6×256 (x2)	
fc6	4096 (x2)	
fc7	4096 (x2)	
L-ATT	1024 (x2)	
fc8	4096 (x2)	
fc9	4096	
f_c	M	

A. Baseline Network Architecture

Similar to the work of [28], the backbone of the SLATT is an AlexNet CNN [56]. Following this architecture, the network consists of five convolution layers and three fully connected layers, where the first two fully connected layers transform features to sparse high-level representations, while the third fully connected layer acts as a classification layer. As part of the AlexNet, convolution layers one, two, and five are followed by max pooling layers to down-sample features and increase robustness to small translations, while Rectified Linear Units (ReLU) are used as activations. To increase the generalization, AlexNet introduced Local Response Normalization (LRN) before the activation and max pooling layers. However, since the introduction of Batch Normalization [57], which has shown to increase model accuracy and reduce training time, the LRN has become deprecated. Therefore, we remove the LRN layers and instead apply batch normalization. Similar to the ResNet architecture [58], we apply batch normalization after each convolution layer, but before ReLU activation. As we consider two modalities, the SLATT contains two identical parallel CNNs; each of these is processing either an RGB or a depth image. To learn modality specific features, weights between these networks are not shared.

The input to the system is an RGB/depth image pair $\{I_m^{RGB}, I_m^D\}, 1 \leq m \leq M$ sampled from the m'th person, where M denotes the total number of persons in the training set. The images are processed by corresponding CNN models, resulting in two global feature vectors $\{x_g^{7,RGB}, x_g^{7,D}\} \in R^{4096}$ from 'fc7' of the SLATT, where the subscript g indicates that the feature is *global*. Next, for each modality, we extract local context features as described in the following.

B. Spatial Attention (S-ATT)

The Spatial Attention (S-ATT) module applies a soft attention mechanism similar to that used for image captioning in [38]. Given an input of size $N \times C \times H \times W$, where N is the batch size, C the number of channels, and $\{H, W\}$ the height and width, respectively, the method works by calculating a local context vector $\hat{x} = \sum_i \alpha_i x_i$, which is the weighted sum of all feature vectors at spatial locations $1 \leq i \leq J | J = HW$. As described in [38], weights α_i can be calculated either hard using a stochastic function or soft using a deterministic function. While the performance between the two variations is largely comparable, the latter is more widely used as it can be easily integrated into the rest of a deep neural network. For a more direct comparison with [28], we consider only soft attention in this work. Soft spatial attention is applied in case of both RGB and depth, although, for simplicity, we neglect RGB and D superscripts in the following description.

In soft attention, weights are calculated from a parametrized score function, which outputs the score between an input feature and a reference vector using weights that are updated along with the rest of the CNN. In case of spatial attention, we define the score function as:

$$e_{l,i} = w_{l,i}^T ReLU(W_{x,l,i} x_{l,i} + W_{l,c} x_c), \tag{1}$$

where $e_{l,i}$ is a scalar representing the score between a vector $x_{l,i}$ from layer $1 \leq l \leq L$ at spatial location i and reference vector x_c. $W_{x,l,i}$ and $W_{l,c}$ are parametrized matrices, while $w_{l,i}$ is a parametrized vector. To take advantage of multiple

modalities, RGB and depth features from 'fc6' of the SLATT are fused, and the resulting RGB-D feature, $x^{6,RGBD} \in R^{2048}$, is used as reference vector in the S-ATT module. To capture correspondences between modalities, features are fused by multiplication to capture higher-order dependencies, and fed to a fully connected layer, resulting in a feature vector consisting of values $x_i^{6,RGBD} = \sum_{j=1}^{2048} w_{ij} x_j^{6,RGB} x_j^{6,D}$, where w_{ij} is learned through back propagation. Thus, spatial attention scores are based on the multimodal behavior of the SLATT.

Weights at each spatial location are calculated by normalizing $e_{l,i}$ using a softmax function, defined as:

$$\alpha_{l,i} = \frac{\exp(e_{l,i})}{\sum_i \exp(e_{l,i})}, \tag{2}$$

where $\alpha_{l,i}$ is the attention weight at spatial location i for the l'th layer. Finally, the local context vector for layer l is calculated as:

$$\hat{x}_l = \sum_i \alpha_{l,i} x_{l,i}, \tag{3}$$

where the length of the context vector depends on the number of feature maps for a particular layer of the network. For our model, the output sizes are provided in Table I. In Section IV-C, we conduct an ablation study, which shows the accuracy by extracting and fusing local context features from S-ATT at different layers of the SLATT.

C. Layer-wise Attention (L-ATT)

The introduction of spatial attention implies that local context features are extracted at different abstraction levels. Still, in the work of [28], these are fused simply by concatenation. By doing so, low-level features containing information about, for example, texture are weighted equally to more high-level features describing larger parts, such as accessories. This is not expedient in case of an input with uniform textures and colors, or where different persons carry accessories that are similar in appearance. Instead, we propose that each local context feature is weighted depending on the input. Thereby, we accomplish a more dynamic fusion scheme, which learns to consider feature importance in relation to the overall accuracy of the system. The dynamic weighting scheme is referred to as layer-wise attention (L-ATT).

As the number of feature maps differs between S-ATT modules, and the L-ATT requires features to be of same size, they are first aligned. This is accomplished by a transformation, $T : R^p \rightarrow R^q$, where p is the size of the feature, i.e., the number of feature maps, while q is the size of the aligned feature. To that end, a linear transformation is applied, which is defined by $\tilde{x}_l = W_l \hat{x}_l + b_l$. In our network, this is implemented using a fully connected layer, where W_l and b_l are the weight and bias, respectively, that are learned along with the rest of the network during training. In Section IV-C, an ablation study is conducted by varying the size of q.

The proposed layer-wise attention module follows an approach similar to that for the S-ATT modules. Given *K* local context feature vectors, weights β_l are calculated from the scores between the features and a reference. Similar to (1), we can define a score function as:

$$a_l = w_l^T ReLU(W_{\tilde{x},l} \tilde{x}_l + W_c x_c), \tag{4}$$

where a_l is the score represented as a scalar, $\tilde{x}_l$ is the aligned local context feature, while w_l, $W_{\tilde{x},l}$, and W_c are parametrized vectors and matrices that influence how the particular feature is weighted. Likewise, weights β_l are calculated by softmax normalization:

$$\beta_l = \frac{\exp(a_l)}{\sum_l \exp(a_l)} \tag{5}$$

Finally, the weighted sum of local context features is calculated as:

$$x_{lo} = \sum_l \beta_l \tilde{x}_l, \tag{6}$$

where the subscript lo indicates the feature containing *local* information.

In this work, the layer-wise attention is applied using an AlexNet architecture as backbone, as described in Section III-A. In principle, the module can be applied to any combinations of features at different abstraction levels and using any network, such as ResNet [58] or GoogLeNet [59].

Multi-level features are learned for each modality by adding a fully connected layer 'fc8', which takes as input the concatenated global and local features, $x_{gl} = [x_{lo}, x_g]$, and outputs a multi-level feature, x_{ml}. The multi-level modality features x_{ml}^{RBG} and x_{ml}^{D} are fused and used as input to an additional fully connected layer, 'fc9', which outputs a multi-level multimodal feature, x_{ml}^{RGBD}, used for classification.

Classification is implemented as a fully connected layer followed by a softmax layer, calculating the probability of a person belonging to the correct class. Given the feature $x_{ml,m}^{RGBD}$, calculated from the input image pair $\{I_m^{RGB}, I_m^{D}\}$, the probability is calculated as: Along with the true label, m, the logistic loss function is used to calculate the error over the entire batch of size N, defined as:

Classification is implemented as a fully connected layer followed by a softmax layer. Given an input pair, $\{I_m^{RGB}, I_m^{D}\}$, the probability of a person belonging to the correct class, given the feature $x_{ml,m}^{RGBD}$, is defined as $\hat{p}_i = Pr(y = m | x_{ml,m}^{RGBD})$. Along with the true label, m, the logistic loss function is used to calculate the error over the entire batch of size N, defined as:

$$L_{id} = -\frac{1}{N} \sum_{i=1}^{N} \log(\hat{p}_i) \tag{7}$$

IV. Experiments

Extensive experiments are conducted on three RGB-D datasets that are all collected from an overhead viewpoint. First, details on training of the SLATT are described in Section IV-A, which is followed by a description of the three evaluated datasets in Section IV-B. Ablation studies are presented in Section IV-C, and the results are used as basis in the experimental results in Section IV-D. A visual analysis is presented in Section IV-E, and the results are finally compared to state-of-the-art systems in Section IV-F.

A. Implementation Details

Training of the SLATT follows a two-step approach. First, modality-based CNNs are trained individually to adapt network parameters to the context of classifying persons within respective domains. In both cases, weights are initialized from a model pre-trained on the ImageNet dataset [56]. Training is performed using Stochastic Gradient Descent (SGD) with a base learning rate of $\eta^0 = 0.001$ and reduced by $\eta^i = \eta^{i-1}0.99$ after each epoch. To further accelerate the training, we add a momentum of $\mu = 0.9$ and train with a batch size of 128. To increase the amount of data and make the network more invariant to translational changes, common augmentation techniques, such as random cropping and flipping, are applied, and the data is shuffled before each new epoch. In case of cropping, images are resized to 256×256 pixels, and cropping values are drawn from a discrete distribution in the interval [0, 29]. To avoid overfitting and increase generalization, dropout is placed after layers 'fc7', 'fc8', and 'fc9' using probability values 0.5, 0.5, and 0.8, respectively. In case of training the depth-based CNN, depth images need to be converted to an appropriate format to take advantage of the pre-trained ImageNet weights. To that end, a JET colormap is applied, which encodes each depth value to an RGB value; red represents objects that are far away, whereas green to blue, represent objects that are close. Applying a JET colormap is fast and has previously shown to outperform other encoding techniques [32]. Weights from the trained RGB- and depth-based CNNs are used to initialize the convolution layers and the first two fully connected layers of the SLATT model. Weights of the remaining layers are initialized using values drawn from a Gaussian distribution with zero mean and a standard deviation of $\sqrt{1/in\_size}$, where in_size refers to the number of input neurons. Hyper-parameters, which are similar to those used to train RGB and depth CNNs, are used to train the SLATT; in both cases, the training runs for 100 epochs. Training is performed on an Nvidia GTX 1080 and takes up to 1.5 hours for modality-based CNNs and up to 4 hours in case of SLATT.

At test time, multimodal features from 'fc9' of the SLATT are extracted from images of persons captured in different camera views. We follow a multi-shot approach and extract features from all images of each person. Features are then summarized by average pooling. Euclidean distance is calculated between all pairs of persons across views and sorted by distance. Thus, shorter distances indicate increased similarity between pairs.

B. Datasets and Protocols

Evaluation of the SLATT is performed on three datasets: *Depth-based Person Identification from Top* (DPI-T) [48], *Top View Person Re-identification* (TVPR) [25], and *Overhead Person Re-identification* (OPR) [39]; the two former are publicly available. To our knowledge, these are the only RGB-D based re-identification datasets collected from an overhead viewpoint.

a) ***DPI-T****:* This dataset consists of 12 persons captured in an average of five appearances in a hallway. An average of 25 sequences are recorded of each person. These are split into 213 training sequences and 249 test sequences. At test time, all test sequences are matched against all training sequences.

b) ***TVPR****:* This dataset contains recordings of 100 persons appearing twice in a hallway; first walking from left to right and then from right to left. Sequences of the first appearance make up the training set, while those of the second appearance constitute the test set. Similar to DPI-T, during tests, all test sequences are matched against all training sequences. For better comparison with [28], we consider the same 94 of the 100 persons, while also doing the evaluation on Region of Interest (ROI) images that are extracted using the You Only Look Once (YOLO) detector [60][1].

c) ***OPR****:* This dataset contains sequences of 64 persons captured in a canteen area. Each person appears twice; when entering the canteen and again when leaving the canteen. In contrast to DPI-T and TVPR, the evaluation of this dataset follows a protocol that is commonly known from RGB-based datasets, such as Market-1501 [9] or CUHK03 [61]. This implies that the data is randomly split into training and test sets, each containing 32 persons. At test time, re-id is performed on the 32 unseen persons. Additionally, 10 random training/test splits are performed, and the average accuracy is calculated across all 10 iterations.

C. Ablation Studies

An ablation study is conducted by configuring the number of considered local context features when only spatial attention is applied. From an empirical study, in [28], only the outputs from S-ATT4 and S-ATT5 are considered. In this work, more extensive experiments are conducted in order to show the impact on accuracy when either adding or removing local context features from additional S-ATT modules.

Table II shows the impact of adding additional local context features at different abstraction levels, starting by only considering the output from only global features and incrementally adding features from S-ATT5 down to S-ATT1. In this case, similar to [28], features are fused by concatenation. Tests are conducted on the datasets presented in Section IV-B and follow the training protocols described in Section IV-A. Contrary to [28], the best results do not only include the outputs from S-ATT4-5, but rather the outputs from S-ATT2-5 or S-ATT3-5. Since [28] does not consider batch normalization, the results are not entirely comparable, but they still provide a good indication of the relevance of feature types across different datasets. In case of OPR, features from S-ATT2 complement additional local and global features, while this is not the case for DPI-T and TVPR, where accuracy is decreasing if additional features from S-ATT1-2 are included. This could be due to the original resolution of the images in OPR, which is higher and thus enables capture of more detailed information at a lower abstraction level. However, overall we see an increase from adding local context features, which shows the benefit from the S-ATT modules.

[1]Annotations and ROI extraction guide provided at: https://github.com/Lejboelle/TVPR_annotations

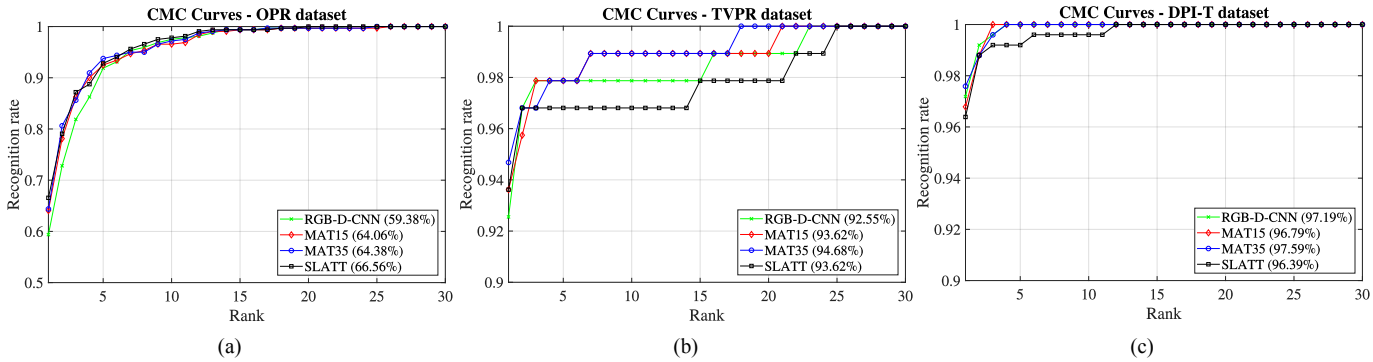

Fig. 3. CMC curves based on experimental results on (a) OPR (p=32), (b) TVPR (p=94), and (c) DPI-T (p=249).

TABLE II
IMPACT ON RANK-1 ACCURACY BY CHANGING THE NUMBER OF S-ATT MODULES IN FUSION OF LOCAL CONTEXT FEATURES. BEST RESULT IN EACH DATASET IS HIGHLIGHTED IN BOLD.

S-ATT1						✓
S-ATT2					✓	✓
S-ATT3				✓	✓	✓
S-ATT4			✓	✓	✓	✓
S-ATT5		✓	✓	✓	✓	✓
OPR Rank-1:	59.38	64.69	63.44	64.38	**65.63**	64.06
DPI-T Rank-1:	94.38	97.19	97.19	**97.59**	97,19	96.79
TVPR Rank-1:	92.55	**94.68**	**94.68**	**94.68**	93.62	93.62

Next, an ablation study is conducted by varying the feature size of the L-ATT module. This impacts both the size of the aligned features, $\tilde{x_l}$, and the size of the output feature, x_{lo}. Table III summarizes the results. In case of DPI-T and TVPR differences are marginal between feature sizes of 256 and 1024, while in case of OPR, a feature size of 1024 increase accuracy by 2.18% and 3.44%, respectively, compared to 512 and 256.

TABLE III
IMPACT ON RANK-1 ACCURACY BY CHANGING THE SIZE OF x_{lo}. BEST RESULT IN EACH DATASET IS HIGHLIGHTED IN BOLD.

	Feature size		
	256	*512*	*1024*
OPR Rank-1:	63.12	64.38	**66.56**
DPI-T Rank-1:	**96.79**	95.98	96.39
TVPR Rank-1:	**94.68**	93.62	93.62

D. Experimental Results

Based on Table III, the following results of the SLATT are based on a feature size of $x_{lo} \in R^{1024}$. Results are presented as Cumulative Matching Characteristic (CMC) curves, that is, for each rank-*i*, a cumulative score is calculated, which represents the percentage of persons having their truth match within the *i* most considered. The results are compared with application of only spatial attention, in this case consideration of S-ATT3-5 (MAT35), as appears from the results in Table II, but also of S-ATT1-5 (MAT15), which provides a more direct comparison when additional layer-wise attention is applied. Furthermore, results are compared with the baseline RGB-D-CNN architecture [39] without attention to show the benefit of fusing global and local information. CMC curves showing accuracies on OPR, TVPR, and DPI-T are shown in Figure 3 (a), (b), and (c), respectively. In case of TVPR, the RGB-D-CNN network is able to re-id almost all persons in the dataset. Since this data was acquired in controlled environmental settings, the only real challenge is the rotational change from walking horizontally in both directions. Improving this result is, therefore, a difficult task. Nonetheless, MAT35 increases the accuracy by 2.13% compared to RGB-D-CNN. Thus, adding local features increases the overall accuracy, although low-level features from S-ATT1 and S-ATT2 do not add additional discriminative information, as also seen in Table II. Nevertheless, it is also worth noting the rank-2 accuracy, which is similar between RGB-D-CNN and MAT35. This indicates the importance of including local features to distinguish between persons with much similar appearance.

A similar result is seen in case of DPI-T, where the results of MAT35 and SLATT are almost identical, where the SLATT, in this case, is inferior to MAT35. Since this dataset consists of only 12 persons, where several sequences are captured of each person, the accuracy of this dataset also seems to be saturated at 97.59% and is therefore difficult to increase. Due to saturated accuracies both on DPI-T and TVPR, we analyze the contribution of the L-ATT module by comparing single-shot and multi-shot accuracies in Subsection IV-G.

The results on the more challenging OPR dataset clearly show the benefit of weighting local features dynamically. While MAT15 shows the smallest increase in accuracy of 4.68% compared to RGB-D-CNN, MAT35 increases the accuracy by 5.00%, while SLATT shows an increase of 7.18%.

E. Visual Attention Analysis

To obtain a better understanding of the relevance of local context features at different abstraction levels, we visualize spatial attention maps from S-ATT modules, which will henceforth be referred to as S-ATT maps, along with their corresponding L-ATT weights. The goal of this analysis is twofold: (1) to identify which local context features are captured at each abstraction level, and (2) to identify trends in the dynamic weighting of features in relation to the dataset. We show examples of success cases to identify discriminative feature regions that result in correct re-identification. Examples are

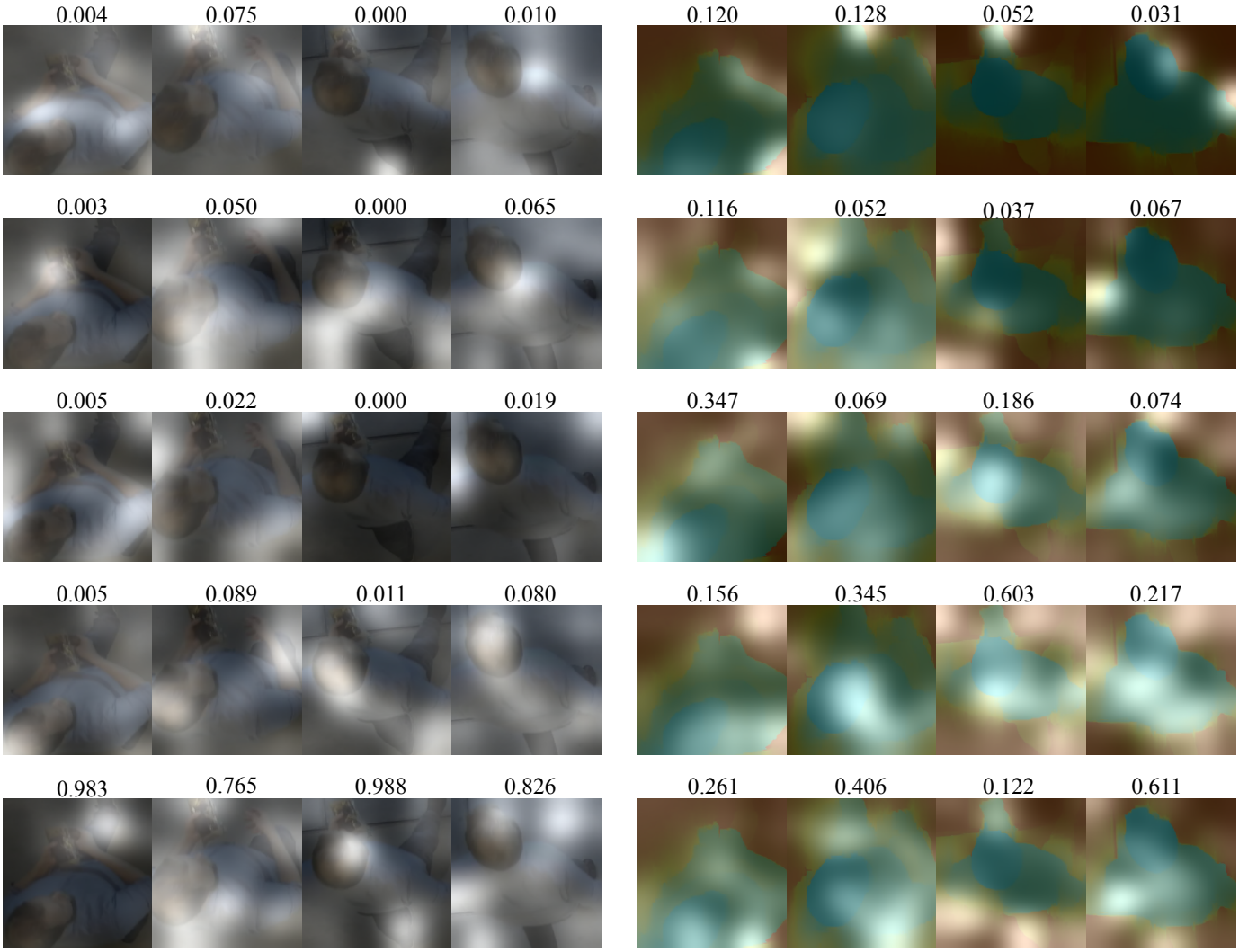

Fig. 4. Visualization of RGB-based (left) and depth-based (right) S-ATT maps with corresponding L-ATT weights for the OPR dataset. Each row shows S-ATT maps from four randomly sampled images of the same person; the first row shows the output from S-ATT1 down to S-ATT5 at the bottom.

shown for all datasets presented in Section IV-B by randomly sampling four images from a person in each dataset and calculating S-ATT maps along with L-ATT weights. Figures 4, 5, and 6 show examples of calculated weights in case of OPR, TVPR, and DPI-T, respectively. Each row shows the S-ATT maps from a single layer, going from S-ATT1 at the top to S-ATT5 at the bottom. RGB-based S-ATT maps are shown to the left, while the depth-based ones are shown to the right. Above the S-ATT maps, layer-wise weights are shown.

The RGB-based S-ATT maps that are shown in case of OPR in Figure 4 indicate a trend to mostly weight the output of S-ATT5, which is the case for all four images. Even though S-ATT5 is highly weighted, differences in S-ATT maps are seen. While the first image captures information around the legs, the second one captures information at the head and shoulder regions, while the third highlights head and legs. More diverse L-ATT weights are seen in case of depth images. S-ATT maps generally tend to highlight regions around the edges, for example at the head/shoulders or around the entire body. While low-level S-ATT maps are mostly concentrated around a few points of interest, S-ATT maps at higher abstraction levels include larger edge regions. A general trend is seen for S-ATT maps, but the L-ATT module is able to dynamically weight features depending on the input, as shown by the differences across the four images. Although weights are distributed more evenly across layers, the outputs of S-ATT3-5 are generally weighted higher.

When the L-ATT weights in Figure 5 are inspected, a trend similar to that in Figure 4 is seen in case of RGB, where features at higher abstraction levels are weighted higher by the L-ATT module. The S-ATT maps show more similarities across the four images, where mostly the head and shoulders are highlighted. Nonetheless, the dynamic weighting causes different information to be fused by weighting low-level features higher in the first image compared to the three other images. Similar to Figure 4, in case of depth, L-ATT weights are more evenly distributed, although features at lower abstraction levels are weighted higher. S-ATT maps are also more centered around few edge points across all layers, while, in case of OPR, this applies typically at lower abstraction

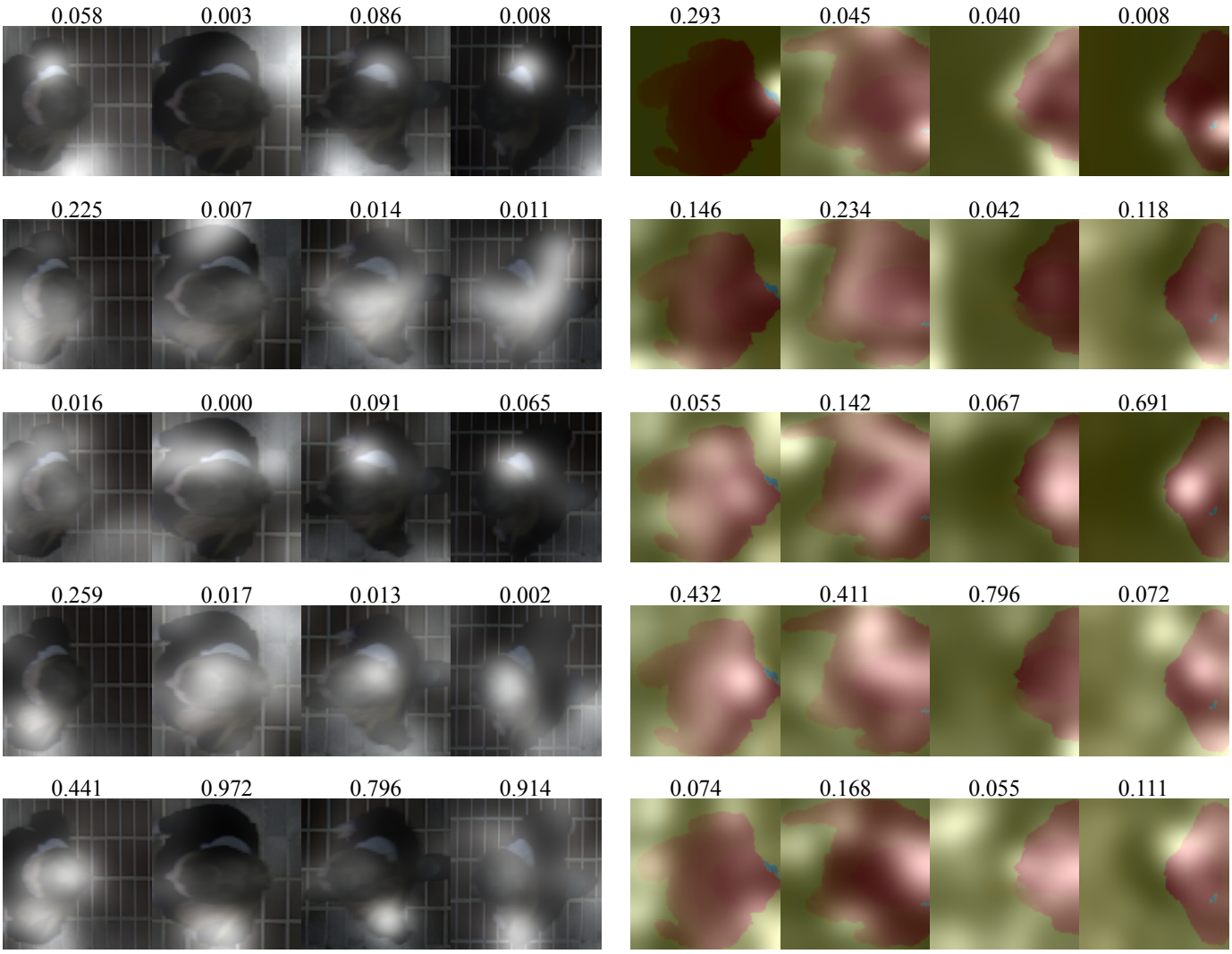

Fig. 5. Visualization of RGB-based (left) and depth-based (right) S-ATT maps with corresponding L-ATT weights for the TVPR dataset. Each row shows S-ATT maps from four randomly sampled images of the same person; the first row shows the output from S-ATT1 down to S-ATT5 at the bottom.

levels.

Larger differences in RGB-based L-ATT weights are seen in Figure 6. The first two images show more evenly distributed weights, whereas the last two mostly weight features at S-ATT5, but they still add complementary low-level information. Larger differences are also seen in case of the S-ATT maps, where both the legs, the frontal body, and the head are highlighted. Similar to OPR and TVPR, depth-based S-ATT maps are centered mostly around edges of the body. However, in contrast to the two former, the head is less highlighted. Likewise, L-ATT does not weight S-ATT1 or S-ATT5 higher, but it distributes weights at all abstraction levels more evenly.

OPR and TVPR that are captured from a more vertical viewpoint and in a less complex scene compared to DPI-T generally weight higher local RGB information at higher abstraction levels, which could be due to less visible texture. This is also indicated by the RGB-based S-ATT maps, which highlight the head and shoulder regions. The depth-based S-ATT maps are more similar across all three datasets as they mostly highlight body edges. Still, while OPR and TVPR place higher weight on the mid- and higher-level features, DPI-T, also in this case, weights low-level features. In all cases, the dynamic weighting scheme ensures fusing of the most relevant features, which are extracted at different abstraction levels. The differences in L-ATT weights across the dataset, which are especially clear when comparing OPR and TVPR to DPI-T, show the strength of the L-ATT to properly weight features at different abstraction levels depending on the data.

Finally, Figure 7 shows cases of incorrect re-id to identify challenging issues in the SLATT. The L-ATT weights show similar trends as for correct re-id. Therefore, the issue lies in the input and the S-ATT maps. In case of both (a) DPI-T and (b) TVPR, RGB-based S-ATT maps are centered around few similar areas: the arm in case of the former, and the hairline in case of the latter. For DPI-T, the depth-based S-ATT maps also mostly highlight the arm, which causes redundant information to be fused. The depth images in (b) also show areas of undefined depth, which is indicated by blue regions, and this results in noisy information. In case of (c) OPR, the S-ATT5 map, which is by far weighted the highest, is quite

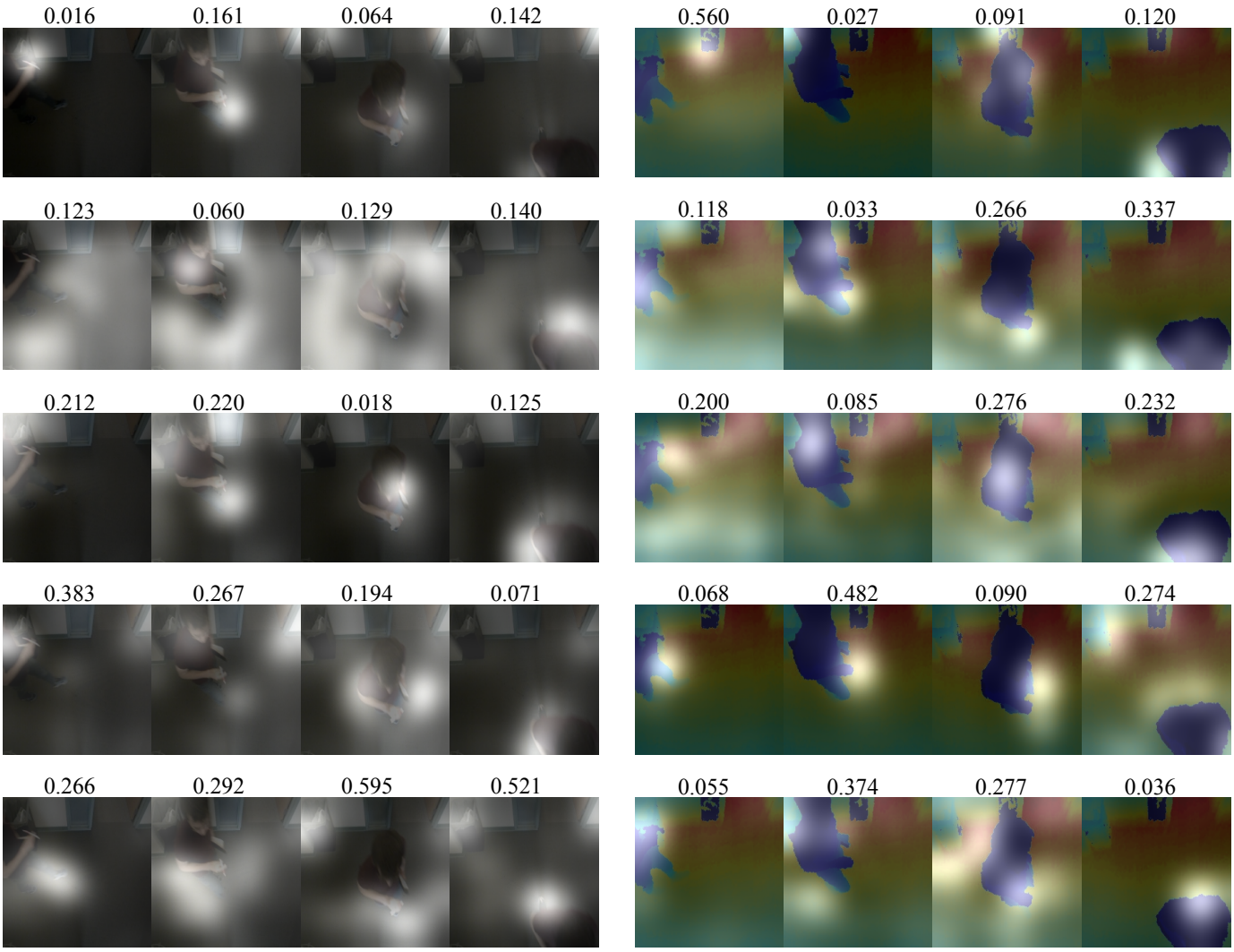

Fig. 6. Visualization of RGB-based (left) and depth-based (right) S-ATT maps with corresponding L-ATT weights for the DPI-T dataset. Each row shows S-ATT maps from four randomly sampled images of the same person; the first row shows the output from S-ATT1 down to S-ATT5 at the bottom.

sparse, and this causes capture of noisy information. Sparsity is also seen in S-ATT4 and S-ATT5 maps in Figure 4. These are, however, different from this failure case as the less noisy S-ATT4 features in Figure 4 are weighted higher. The depth-based S-ATT maps highlight more non-relevant areas, such as the plate or the floor. This is especially seen when inspecting S-ATT3 and S-ATT4 maps. This could indicate difficulties when a person carries objects that are common to the scene, in this case a plate of food.

F. Comparison with State-of-the-Art Systems

Comparisons between the results of the SLATT, presented in Figure 3, and state-of-the-art systems are provided in Tables IV-VI.

Previously proposed systems have evaluated the DPI-T dataset using only depth information. As a result, we compare the results of our SLATT and previous RGB-D CNNs by extracting depth features, which is indicated by the subscript D. We compare the results with the residual attention (4D RAM) proposed in [48] and with CNN-LSTM (Depth ReID) proposed in [37] along with the RGB-D-CNN [39] and MAT [28], both with and without the use of batch normalization. MAT35$_D$ (ours) refers to the results of the MAT, which considers additional local context features from S-ATT3. Furthermore, we also provide comparisons of RGB-D-CNN, MAT, MAT35, and SLATT with RGB information included. In all cases, MAT35 and SLATT make use of batch normalization.

As seen in Table IV, the use of batch normalization clearly increases the accuracy, which is shown for both baseline RGB-D-CNN and MAT. Moreover, including additional local information at lower abstraction levels decreases the accuracy when comparing MAT$_D$+BN and MAT35$_D$. This could indicate that low-level depth features do not provide enough discriminative information to ensure benefits. This could also be the reason why the SLATT$_D$ provide a rank-1 accuracy which is inferior to MAT$_D$+BN. However, the addition of RGB information increases the accuracy by up to 14.46% when comparing MAT$_D$+BN and MAT35 while SLATT provide accuracies almost similar to MAT35. Even though the accuracy is high when including only depth information, the complementarity

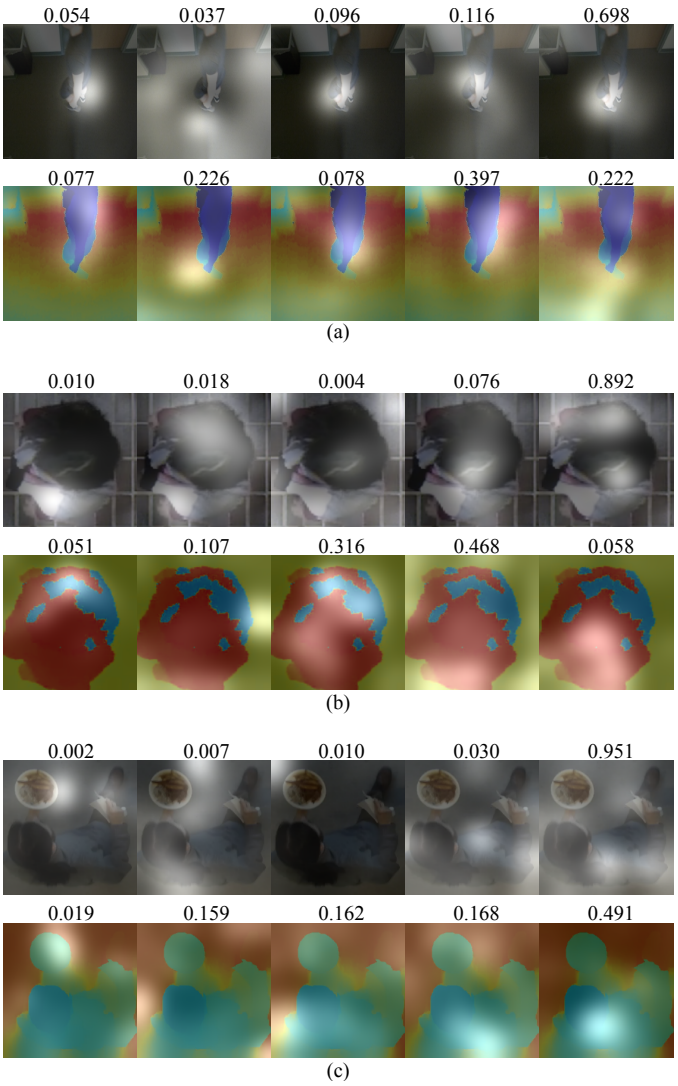

Fig. 7. Visualization of S-ATT maps with corresponding L-ATT weights in failure cases for (a) DPI-T, (b) TVPR, and (c) OPR. Both RGB-based S-ATT maps (top) and depth-based attention maps (bottom) are shown from the S-ATT1 (left) to S-ATT5 (right).

of RGB and depth, combined with the use of both local and global features, produces more discriminative features, which results in higher accuracy.

Besides RGB-D-CNN and MAT, with and without batch normalization, the only other comparable system in case of TVPR, as seen in Table V, is the one of [25], where hand-crafted RGB-D features are extracted (TVDH). Similar to DPI-T, the addition of batch normalization results provides a significant increase in accuracy, while CNN-based features outperform the hand-crafted ones by up to 19.38% when comparing TVDH and MAT35. In contrast to DPI-T, additional information from S-ATT3 does not increase the accuracies when comparing MAT+BN and MAT35. When adding layer-wise attention, we do not benefit from additional low-level information and achieve a rank-1 accuracy similar to that of MAT35. This could be due to accuracy being close to saturated or the resolution of depth, which result in uniformly colored images after applying the JET color map.

TABLE IV
COMPARISON BETWEEN SLATT AND STATE-OF-THE-ART SYSTEMS ON THE DPI-T DATASET (P=249) ('–' INDICATE THAT A RESULT IS NOT AVAILABLE). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method/Rank	r = 1	r = 5	r = 10	r = 20
4D RAM [48]	55.60	–	–	–
Depth ReID [37]	76.30	–	–	–
RGB-D-CNN$_D$ [39]	53.82	87.95	99.20	100
RGB-D-CNN$_D$ [39]+BN	80.72	97.59	100	100
MAT$_D$ [28]	53.41	89.16	99.20	100
MAT$_D$ [28]+BN	**83.13**	97.19	100	100
MAT35$_D$ (ours)	81.93	97.99	100	100
SLATT$_D$ (ours)	79.52	97.59	100	100
RGB-D-CNN [39]+BN	94.38	99.20	100	100
MAT [28]+BN	97.19	100	100	100
MAT35 (ours)	**97.59**	100	100	100
SLATT (ours)	96.39	99.20	99.60	100

TABLE V
COMPARISON BETWEEN SLATT AND STATE-OF-THE-ART SYSTEMS ON THE TVPR DATASET (P=94). BEST RESULTS ARE HIGHLIGHTED IN BOLD. (*RESULTS ARE ESTIMATED FROM THE CMC CURVE.

Method/Rank	r = 1	r = 5	r = 10	r = 20
TVDH* [25]	75.50	87.50	89.20	91.90
RGB-D-CNN [39]	80.85	92.55	92.55	95.74
RGB-D-CNN [39]+BN	92.55	97.87	97.87	100
MAT [28]	82.98	93.62	94.68	96.81
MAT [28]+BN	94.68	97.87	97.87	97.87
MAT35 (ours)	**94.68**	97.87	97.87	100
SLATT (ours)	93.62	96.81	97.87	100

Comparisons between SLATT and state-of-the-art systems on the OPR dataset, which is provided in Table VI, indicate that more importance should be directed towards dynamic feature weighting schemes when difficult datasets are being evaluated. As also seen in Table II, adding local features from S-ATT3 increases the accuracy by 0.94% when comparing MAT+BN and MAT35. The rank-1 accuracy is decreased to 64.06% when adding additional local features from S-ATT1 and S-ATT2, as shown in Table II, but dynamically weighting the features using layer-wise attention increases the accuracy by 2.50%. Additionally, compared to the previous work of [28] with BN, the rank-1 accuracy of the SLATT is increased by 3.12% while the accuracy is increased by 7.18% compared to RGB-D-CNN+BN.

To further highlight the significance of the proposed system, we provide pairwise statistics of the rank-1 accuracy on OPR between SLATT and the three systems of MAT, MAT35 and RGB-D-CNN. A comparison is provided as a box plot in Figure 8. From here, it is clear that the variety of MAT is

TABLE VI
COMPARISON BETWEEN SLATT AND STATE-OF-THE-ART SYSTEMS ON THE OPR DATASET (P=32). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method/Rank	r = 1	r = 5	r = 10	r = 20
RGB-D-CNN [39]	45.63	82.81	94.69	99.69
RGB-D-CNN [39]+BN	59.38	91.88	97.50	99.69
MAT [28]	49.06	89.06	95.62	99.38
MAT [28]+BN	63.44	92.50	96.25	99.69
MAT35 (ours)	64.38	93.75	97.19	99.69
SLATT (ours)	**66.56**	92.81	**97.81**	**100**

lower than that of SLATT, however, the maximum observed value of SLATT is higher while the minimum observed value is higher than all three compared systems. Additionally, while the medians of MAT and MAT35 are higher than that of RGB-D-CNN, the median of SLATT is higher than all three.

In addition to Figure 8, we also provide paired t-tests to show the significance in terms of probabilities. That is, between SLATT and the remaining three systems we calculate *t*-values using the differences in rank-1 accuracy between methods during all 10 test runs. Given the t-value, we use a look-up table to infer a corresponding *p*-value, which is an indicator of the level of significance. The t-value is calculate as:

$$t = \frac{\bar{d}}{SE(\bar{d})}, \tag{8}$$

where $\bar{d}$ is the mean of differences while $SE(\bar{d})$ is the standard error of the mean differences, calculated as $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$, where s_d is the standard deviation of differences and n is the number of test iterations, i.e. 10 in our case.

Table VII provides an overview of pairwise t-values and corresponding p-values.

Observing the p-values of SLATT/MAT and SLATT/RGB-D-CNN in Table VII, there is strong evidence that the inclusion of the L-ATT module results in higher accuracies since, in both cases, the value is less than 0.05. Compared to MAT35, the results are marginally significant since the p-value is just above 0.05, which still indicates good evidence of a positive impact on accuracy.

TABLE VII
P- AND T-VALUES FROM PAIRWISE T-TESTS BETWEEN SLATT AND MAT, MAT35 AND RGB-D-CNN, RESPECTIVELY.

	SLATT/MAT35	SLATT/MAT	SLATT/RGB-D-CNN
t-value	2.091	3.001	3.977
p-value	0.066	0.015	0.003

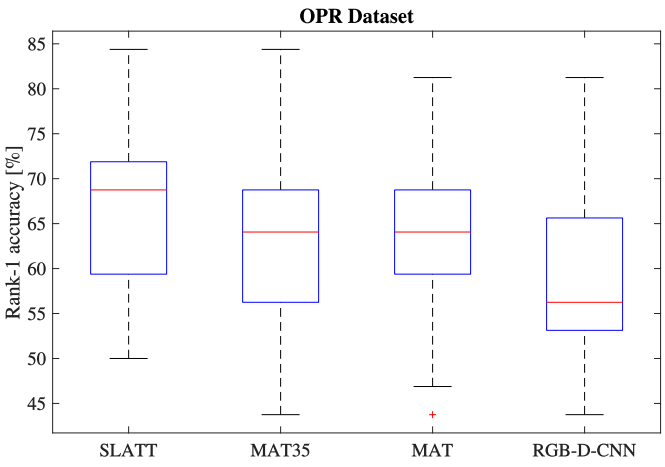

Fig. 8. Comparison of statistical differences between SLATT, MAT35, MAT and RGB-D-CNN.

G. Contribution of L-ATT

From the visual analysis in Section IV-E, it is clear that the L-ATT module is able to dynamically weight features at different abstraction levels based on the input. To further study the effect of this property, we compare the results of the multi-shot setting with a single-shot setting, where only a single image of each person is available at the time of testing. In the single-shot setting, we randomly sample an image from each person in both camera views and, similarly to the multi-shot setting, calculate Euclidean distances between extracted features. In both settings, we consider only the rank-1 accuracy and compare the relative increase from single- to multi-shot accuracy across RGB-D-CNN+BN, MAT+BN, MAT35 and SLATT. Table VIII provides an overview of rank-1 accuracies in case of single- and multi-shot settings, respectively, while Figure 9 shows the relative increase between the two settings.

From Figure 9 it is clear that the addition of the L-ATT module results in an architecture that better captures the individual structures in each image, resulting in an overall larger increase in rank-1 accuracy when fusing features from multiple images. In case of TVPR and DPI-T, the relative increase compared to MAT35 is 8.51% and 0.8%, respectively. Only in the case of OPR do we see similar relative increase when comparing MAT35 and SLATT, however, compared to MAT, the relative increase of SLATT is 3.75% higher.

Interestingly is also the fact that the relative increase of RGB-D-CNN in case of DPI-T and OPR is 1.21% and 13.13%, which is 6.02% and 10.62% worse, respectively, compared to SLATT. This indicates the importance of both capturing local context features using the S-ATT module, and dynamically fuse the features using the L-ATT module.

To further highlight the contribution of the L-ATT module in a setting where identifying the optimal combination of local context features takes much longer, we conduct experiments using a different, deeper, CNN as backbone. We choose an architecture which is comparable to the AlexNet in terms of complexity, to make the results more comparable to those shown in Figure 3. Due to its high performance compared to complexity [62], we choose MobileNetV2 [63] as backbone. The network consists of *bottleneck* operators that each consist of up to four identical *bottleneck residual blocks*, where the number of parameters of the layers depend on the bottleneck. The residual blocks each consist of an expansion layer transforming the input from size H×W×C to H×W×kC by 1×1 convolutions, a depthwise 3×3 convolution layer transforming the input from size H×W×kC to H/s×W/s×kC, and a linear layer transforming the output from the depthwise convolution to size H/s×W/s×C' by 1×1 convolutions. As activation, they use ReLU6, which is a ReLU activation function with an upper bounded value of six. The network consists of seven bottleneck operators, thus, the number of possible combinations of local context features exceeds 5000. Since it is inexpedient to evaluate such a high number of combinations, we compare the result of concatenating the local context features of all seven bottleneck operators to weighting the features using the L-ATT module.

We train RGB and depth CNNs as described in Section

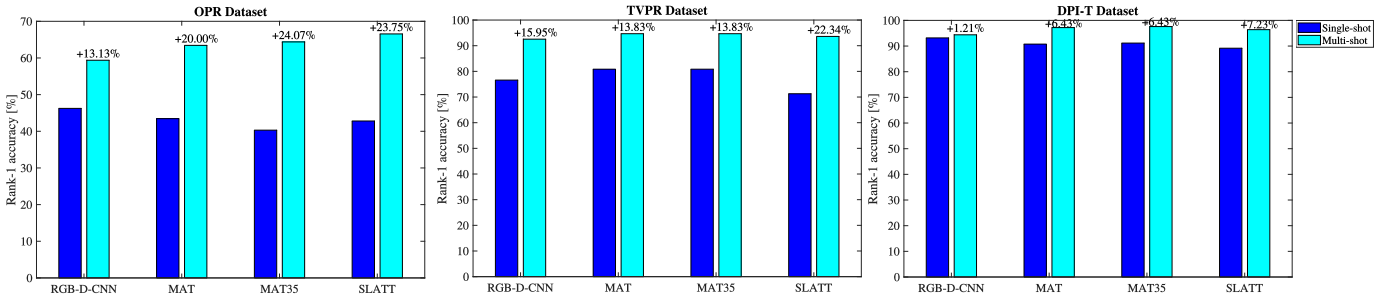

Fig. 9. Relative increase in rank-1 accuracy from single-shot to multi-shot setting using RGB-D-CNN, MAT, MAT35 and SLATT, respectively, on OPR (p=32), TVPR (p=94), and DPI-T (P=249).

TABLE VIII
OVERVIEW SINGLE- AND MULTI-SHOT RANK-1 ACCURACIES ON OPR (P=32), TVPR (P=94) AND DPI-T (P=249) DATASETS.

	OPR		TVPR		DPI-T	
Method/Rank-1	Single-shot	Multi-shot	Single-shot	Multi-shot	Single-shot	Multi-shot
RGB-D-CNN [39]+BN	46.25	59.38	76.60	92.55	93.17	94.38
MAT [28]+BN	43.44	63.44	80.85	94.68	90.76	97.19
MAT35 (ours)	43.44	64.38	80.85	94.68	91.16	97.59
SLATT (ours)	42.81	66.56	71.28	93.62	89.16	96.39

IV-A, and afterwards train SLATT and MAT models, respectively. experiments are conducted on TVPR and OPR[2], following the protocols described in Section IV-B. Training and testing the MAT using a 1080 GTX takes ≈4 hours, thus, it would take a long time to find the optimal set of local features using exhaustive search. The experimental results are shown in Figure 10. On OPR, concatenating features results in a rank-1 accuracy of 70.62%, while the use of dynamic fusion increases rank-1 accuracy by 2.19% to 72.81%. Similarly on TVPR, rank-1 accuracy is increased by 6.38% from a 88.30% to 94.68%. Finally, using MobileNetV2 as backbone in SLATT, rank-1 accuracy is increased by 6.25% and 1.06% on OPR and TVPR, respectively, compared to using AlexNet. From the results it is clear that the proposed SLATT better captures the importance of different local features, while neglecting redundant ones. As a result, only the most informative features are considered, resulting in a higher accuracy.

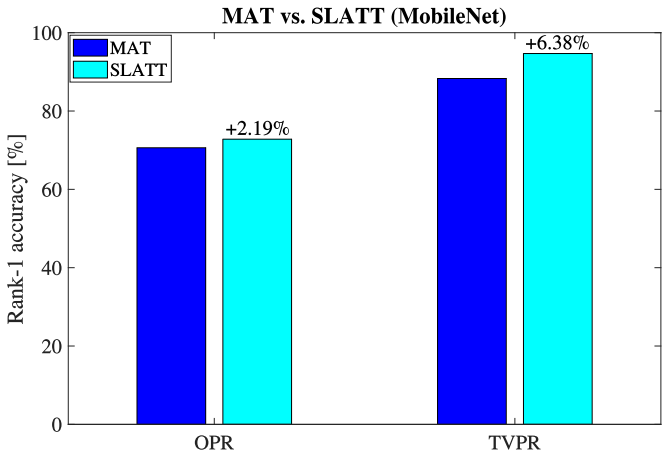

Fig. 10. Comparison of rank-1 accuracy between concatenation of local features (MAT) and dynamic feature fusion (SLATT) using MobileNetV2 as backbone.

V. CONCLUSION

In this work, we combine the use of spatial attention (S-ATT) to capture features at different abstraction levels in a multimodal CNN with dynamic fusion of local context features at different abstraction levels. This is done by introducing a layer-wise attention module (L-ATT), which dynamically weights features based on the input and the multimodal behavior of the entire model. Layer-wise weights are calculated using a soft attention mechanism, which calculates the scores between each of the local context features from the S-ATT modules and a multimodal reference vector, to determine the relevance of each feature. Thus, a weighted summary of features at all abstraction levels makes up a multi-local feature vector containing local discriminative information. Local and global features are fused in case of both RGB and depth, and a multi-level multimodal feature is finally generated by fusion of modality-based features. Experimental results on two public datasets, DPI-T and TVPR, show rank-1 accuracies of 96.39% and 93.62%, respectively, which are comparable to the existing state-of-the-art systems. Additionally, the state-of-the-art accuracy on a third dataset, OPR, is increased by 3.12% compared to previous work. From a visual analysis of both S-ATT maps and corresponding L-ATT weights, it is shown that the L-ATT module is able to adapt the dynamic weighting to the data. Our results on the datasets OPR and TVPR, which are captured from a more vertical viewpoint, show that head and shoulder regions are highlighted and weighted higher compared to DPI-T. Finally, a quantitative analysis highlights the contribution of the L-ATT module by providing higher

[2]Using RGB alone resulted in a rank-1 accuracy of 99.60% on DPI-T, thus, it does not make sense to do further testing of SLATT or MAT on this dataset.

relative accuracies when fusing information from multiple images compared to considering a single image. Additionally, using a deeper CNN as backbone, such as MobileNetV2 that consists of several more local context features, dynamic fusion of features results in a higher rank-1 accuracy on both OPR and TVPR.

VI. Discussion and Future Work

Based on the experimental results and visual analysis, a clear advantage of the proposed system compared to previous work is its ability to capture useful local context information using the S-ATT, which increases the accuracy as also shown in Table II. Additionally, the L-ATT module does not follow a common weighting scheme for all datasets, but adapts to the presented data. Furthermore, it is able to determine the relevance of local context features to the overall multi-modal fusion scheme based on each individual image. This adds a certain robustness to translational and rotational changes. However, challenges arise when the viewpoint becomes more vertical or objects common to the scene are present. The first issue is indicated by the less vertically captured DPI-T dataset, where also more background information is present compared to OPR and TVPR. In this case, discriminative depth information is difficult to exploit since depth maps are more similar across the entire dataset. In this case, a better solution might be to apply a joint localization algorithm, as in [21], to capture relations between body parts. Furthermore, even though, the addition of the L-ATT module show larger minimum, maximum and median rank-1 accuracies on the OPR dataset, more work still needs to be done, to make the method less sensitive to different data distributions in order to minimize variety between tests. In this case, we observed smaller variation in case of MAT. One idea is to also apply dynamic weighting to each frame, as proposed in [37], to suppress, or even neglect, noisy frames.

References

[1] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person re-identification*. Springer, 2014, pp. 1–20.

[2] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[4] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, preprint.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015, pp. 2197–2206.

[6] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proc. CVPR*, 2016, pp. 1363–1372.

[7] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Proc. AAAI*, 2016, pp. 3655–3661.

[8] F. M. Khan and F. Brèmond, "Multi-shot person re-identification using part appearance mixture," in *Proc. WACV*, 2017, pp. 605–614.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015, pp. 1116–1124.

[10] S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2963–2977, 2018.

[11] K. Li, Z. Ding, S. Li, and Y. Fu, "Toward resolution-invariant person reidentification via projective dictionary learning," *IEEE transactions on neural networks and learning systems*, 2018, (Early Access).

[12] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, 2012, pp. 2288–2295.

[13] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. CVPR*, 2016, pp. 1239–1248.

[14] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. CVPR*, 2017, pp. 2530–2539.

[15] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 791–805, 2018.

[16] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, 2015, pp. 3908–3916.

[17] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proc. CVPR*, 2016, pp. 1335–1344.

[18] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proc. AVSS*, 2017, pp. 1–6.

[19] S. Wu, Y.-C. Chen, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. WACV*, 2016, pp. 1–8.

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, 2014, pp. 34–39.

[21] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. CVPR*, 2017, pp. 1077–1085.

[22] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.

[23] M. Gou, S. Karanam, W. Liu, O. I. Camps, and R. J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset." in *Proc. CVPR Workshops*, 2017, pp. 1425–1434.

[24] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person trasfer gan to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018, pp. 79–88.

[25] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti, "Person re-identification dataset with rgb-d camera in a top-view configuration," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. Springer, 2016, pp. 1–11.

[26] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, 2018, pp. 420–429.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[28] A. R. Lejbølle, B. Krogh, K. Nasrollahi, and T. B. Moeslund, "Attention in multimodal neural networks for person re-identification," in *Proc. CVPR Workshops*, 2018.

[29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[30] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. CVPR*, 2017, pp. 384–393.

[31] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, 2018, pp. 402–419.

[32] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *Proc. IROS*, 2015, pp. 681–687.

[33] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," in *Proc. ICRA*, 2015, pp. 1329–1335.

[34] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks," in *Proc. CVPR*, 2017, pp. 416–425.

[35] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *Proc. ICRA*, 2017, pp. 4628–4635.

[36] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2588–2603, 2017.

[37] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto, "Reinforced temporal attention and split-rate transfer for depth-based person re-identification," in *Proc. ECCV*, 2018, pp. 715–733.

[38] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[39] A. R. Lejbølle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Multimodal neural network for overhead person re-identification," *Proc. BIOSIG*, pp. 25–34, 2017.

[40] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, 2017, pp. 3219–3228.

[41] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[42] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. CVPR*, 2018, pp. 1179–1188.

[43] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, 2018, pp. 5363–5372.

[44] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294.

[45] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.

[46] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. CVPR*, 2017, pp. 3156–3164.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[48] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proc. CVPR*, 2016, pp. 1229–1238.

[49] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. CVPR*, 2018, pp. 369–378.

[50] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Task-driven dynamic fusion: Reducing ambiguity in video description," in *Proc. CVPR*, 2017, pp. 3713–3721.

[51] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. ICCV*, 2017, pp. 4203–4212.

[52] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *Proc. AAAI*. 7202–7209, 2018.

[53] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, 2017, pp. 1831–1840.

[54] H. Chen, Y.-F. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for rgb-d salient object detection," in *Proc. IROS*, 2018, pp. 6821–6826.

[55] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. CVPR*, 2018, pp. 2109–2118.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[57] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[60] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. CVPR*, 2017, pp. 6517–6525.

[61] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014, pp. 152–159.

[62] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018.

[63] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.