STNReID : Deep Convolutional Networks with Pairwise Spatial Transformer Networks for Partial Person Re-identification

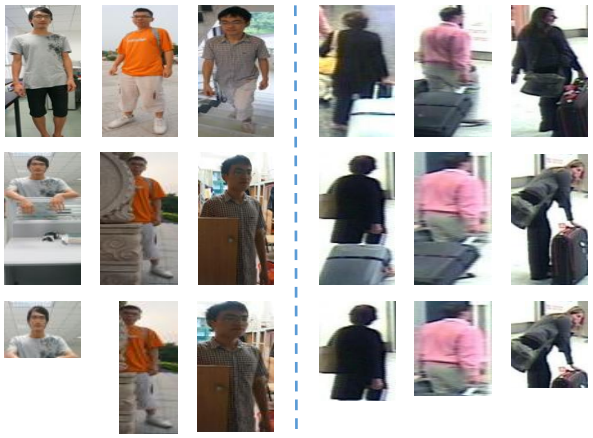
Hao Luo[1,2] , **Xing Fan**[1,2] , **Chi Zhang**[2] and **Wei Jiang**[1]
[1]Zhejiang University
[2]Megvii Inc
{haoluocsc, xfanplus, jiangwei_zju}@zju.edu.cn, zhangchi@megvii.com

Abstract

Partial person re-identification (ReID) is a challenging task because only partial information of person images is available for matching target persons. Few studies, especially on deep learning, have focused on matching partial person images with holistic person images. This study presents a novel deep partial ReID framework based on pairwise spatial transformer networks (STNReID), which can be trained on existing holistic person datasets. STNReID includes a spatial transformer network (STN) module and a ReID module. The STN module samples an affined image (a semantically corresponding patch) from the holistic image to match the partial image. The ReID module extracts the features of the holistic, partial, and affined images. Competition (or confrontation) is observed between the STN module and the ReID module, and two-stage training is applied to acquire a strong STNReID for partial ReID. Experimental results show that our STNReID obtains 66.7% and 54.6% rank-1 accuracies on partial ReID and partial iLIDS datasets, respectively. These values are at par with those obtained with state-of-the-art methods.

1 Introduction

Person re-identification (ReID), especially on holistic person datasets, has achieved huge improvements using deep learning techniques in recent years [Sun *et al.*, 2018; Zhang *et al.*, 2017; Wei *et al.*, 2017]. However, partial observations of person images exist due to occlusions, viewpoints, and special tasks in real-world applications. For example, a target person inside a subway station, airport, or supermarket may be occluded by ticket gates, baggage and checkout counters, or other things/people. The patch of a person's body is constantly maintained by using several techniques, such as human detection or skeleton model, to reduce such information interference. In such situations, most ReID models trained on holistic person datasets are unstable and frequently fail in searching for accurate images. Hence, Zheng *et al.* [Zheng *et al.*, 2015b] addressed the partial person re-identification (partial ReID) task and proposed the Partial-ReID dataset. In recent years, the partial ReID has gradually attracted the attention of both researchers and engineers due to its huge application value.



(a) Partial-ReID (b) Partial-iLIDS

Figure 1: Example of holistic person images (first row), annotated partial person images for recognition (third row), and the corresponding non-partial images (second row).

Partial ReID is a challenging task because only partial information is available for matching target persons (as shown in Fig. 1). It is also made difficult by the need to determine whether a partial image should be resized to a fix-sized image. Undesired deformation exists when the model is resized to a fix-sized image. By contrast, the model should match an arbitrary image patch to a differently sized holistic image when the partial image is not resized. Two features or feature maps with different sizes cannot be compared, especially on end-to-end convolutional networks (CNNs). In [He *et al.*, 2018a], He *et al.* summarized the drawbacks of several kinds of solutions. For example, Sliding Window Matching (SWM) [Zheng *et al.*, 2015b] establishes a sliding window with the same size as the partial image and utilizes it to search for the most similar region on each holistic image. However, the computational cost of this solution is expensive due to the traversal calculation. Deep Spatial Reconstruction (DSR) [He *et al.*, 2018a] directly matches two arbitrary-sized feature maps by sparse reconstruction in a deep learning frame-

work to accelerate the matching process. Although DSR is an impressive work for partial ReID, its arbitrary-sized matching mechanism limits the utilization of the tensor computing power of the GPU device. DSR requires one-to-one matching after the extraction of the feature maps of partial and holistic images. It also becomes inefficient with the enlargement of image size. In CNNs, a batch is composed of several tensors (images) with the same sizes. Specifically, CNNs should deal with "resizing" and "matching" problems at the same time.

In the present study, we present a novel deep ReID framework (STNReID) based on spatial transformer networks [Jaderberg *et al.*, 2015] for partial ReID. STNReID includes one spatial transformer network (STN) module and one ReID module. The STN module utilizes partial and holistic images and predicts the parameters of 2D affine transformers, such as resizing (scaling), rotation, and reflection, of two images. Then, the STN module samples an affined image from the holistic image to match the partial image. The ReID module extracts the features of the affined and partial images to retrieve the target person images. Trained partial images are generated from holistic person datasets. Hence, STNReID can be trained without the need for additional labeled data. However, the performance of the STN module decreases with the increase of the power of the ReID module. Therefore, a weak ReID is used to train a strong STN module, and the STN module is restricted to fine-tune the ReID module. Such two-stage training enables the acquisition of a strong STNReID model. In the inference stage, STNReID can match a certain partial image with several holistic images in a batch. This end-to-end one-to-many matching is more efficient than one-to-one matching is. The major contributions of our work are summarized as follows:

- A novel partial ReID framework (STNReID) based on STNs is proposed. The proposed STNReID can be trained on holistic person datasets without the need for additional labeled partial images. To our best knowledge, this pairwise STNs is involved into person ReID task at the first time.
- We found the STNs can be trained by the semantic features and its performance is affected by the ReID module. A two-stage training process is proposed to enable the acquisition of a strong STNReID for partial ReID.
- The experimental results show that our STNReID achieves competitive outcomes on Partial-ReID [Zheng *et al.*, 2015b] and Partial-iLIDs [Zheng *et al.*, 2011] benchmarks. And the better baseline will benefit the research community.

2 Related Works

In this section, deep learning-based person ReID methods are summarized, and the existing relevant studies on partial ReID are reviewed because partial ReID is a sub-topic of person ReID. Then, STNs [Jaderberg *et al.*, 2015] and their application to person ReID are investigated.

2.1 Deep person ReID

Deep learning-based person ReID uses deep CNNs (DCNNs) to represent the features of person images. On the basis of the loss functions of training DCNNs, most existing studies focus on two methods, namely, robust representation learning and deep metric learning. Representation learning-based methods [Zheng *et al.*, 2016; Zheng *et al.*, 2017] aim to learn robust features for person ReID by using the Softmax loss (ID loss). An ID embedding network (IDENet) [Zheng *et al.*, 2016; Zheng *et al.*, 2017] regards each person ID as a category of a given classification problem. In addition, Fan *et at.* [Fan *et al.*, 2019] obtained the variants of the SoftMax function and achieved superior performance in the field of ReID.

Compared with representation learning, deep metric learning-based algorithms directly learn the distance of an image pair in the feature embedding space. The typical metric learning method is the triplet loss [Liu *et al.*, 2017], which pulls the distance of a positive pair and pushes the distance of a negative pair. However, triplet loss is easily influenced by the selected samples. Hard mining techniques [Hermans *et al.*, 2017; Xiao *et al.*, 2017; Ristani and Tomasi, 2018] widely used to obtain triplet loss with high accuracies. Improved triplet loss [Cheng *et al.*, 2016] and quadruplet loss [Chen *et al.*, 2017] are variants of the original triplet loss. At present, the combination of ID loss with triplet loss has attracted considerable attention due to its remarkable performance.

2.2 Partial ReID

With the development of person ReID, partial ReID has gradually attracted the attention of researchers due to its huge value in real-world ReID applications. However, few studies have solved the task, and the performance of partial ReID is unsuitable for practical applications. As a solution to this problem, many methods [Donahue *et al.*, 2014; Girshick *et al.*, 2014] have been designed such that they are able to directly resize an arbitrary patch of a person image to a fixed-size image and extract the fixed-length global features for matching. However, the various scale deformations caused by such rough methods are difficult to address. In this regard, part-based models provide an optional solution. In [Zheng *et al.*, 2015b], Zheng *et al.* proposed a global-to-local matching model called SWM that can capture the spatial layout information of local patches and introduced a local patch-level matching method called the Ambiguity-sensitive Matching Classifier (AMC) that is based on a sparse representation classification formulation with an explicit patch ambiguity model. However, the computation cost of the AMC–SWM is extremely expensive because the extraction of features requires considerable time without sharing computation. Apart from handcrafted features or models, He *et al.* [He *et al.*, 2018a] used a fully convolutional network to generate spatial feature maps with certain sizes and whose pixel-level features are consistent. Then, DSR is conducted to match a pair of person images with different sizes.

2.3 Spatial Transformer Networks

Spatial Transformer Networks (STNs), which include a localization network and a grid generator, make up the deep learning method proposed in [Jaderberg *et al.*, 2015]. The localization network utilizes feature maps and outputs the parameters of 2D affine transformations through several hidden

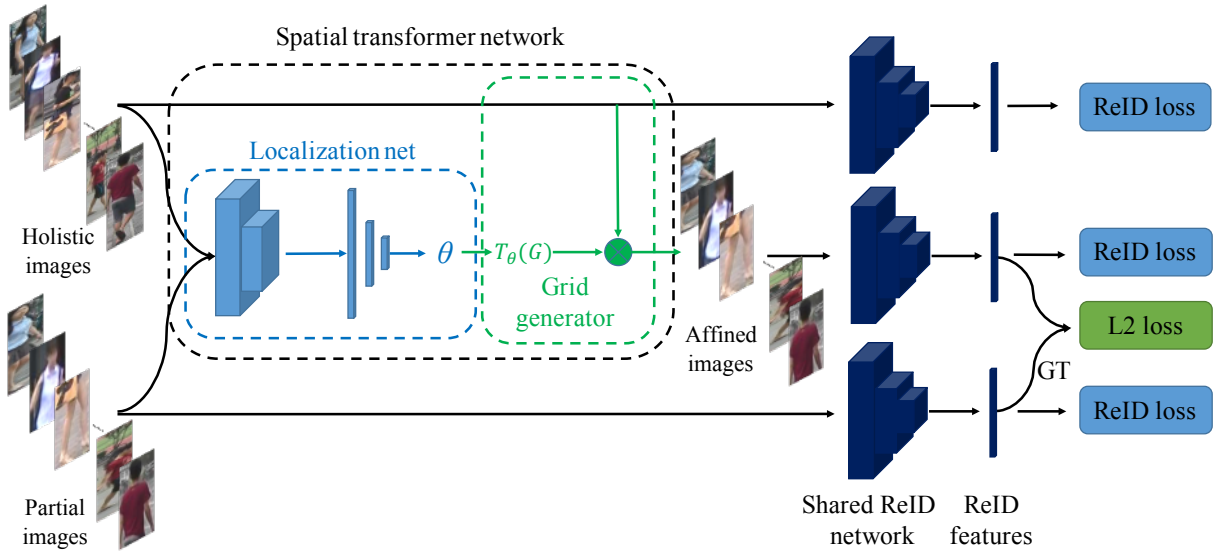


Figure 2: The framework of STNReID, which includes an STN and a ReID module. The ReID loss combines ID loss and Triplet loss.

layers. Then, such predicted transformation parameters are transferred in the grid generator to create a sampling grid, which is a set of points where the input feature map is sampled to produce the transformed output. An STN can perform 2D affine transformations, such as reflection, rotation, scaling, and translation. For person ReID, Zheng *et al.* [Zheng *et al.*, 2018] propose Pedestrian Alignment Network (PAN), which combines the STN and a deep ReID network. However, PAN is similar to STNs as it predicts parameters on the basis of the feature maps of one image and aligns the weak spatial changes of holistic person images.

3 Methods

This section introduces the architecture of the proposed STNReID and demonstrates the training of an improved STNReID in a two-stage pipeline.

3.1 STNReID

As shown as Fig. 2, STNReID includes an STN module and a ReID module. The STN module predicts the parameters of the affine transformations, and then crops patches (affined images) from holistic person images to match partial images. The ReID module extracts the global features of holistic images, partial images and affined images.

STN Module

The holistic and partial images are denoted as I_h and I_p, respectively. In the training stage, the partial image is randomly cropped from the holistic image. I_h and I_p are the tensors with the shape of $H \times W \times C$, where H and W the height and width of the resized images with fixed sizes, respectively; and C is the number of image channel (e.g. $C = 3$ for RGB images). Then, I_h and I_p are concatenated as a new tensor $I_{h,p}(H \times W \times 2C)$ and is fed in the localization network of the STN module. The localization network outputs θthrough a series of hidden layers, and the parameters of transformation $T(\theta)$: $\theta = f_{loc}(I_{h,p})$. The localization network function $f_{loc}()$ is dependent on the structure of the localization network.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where $(x_i^t, y_i^t)^T$ are the target coordinates of the regular grid in the affined image, $(x_i^s, y_i^s)^T$ are the source coordinates in the holistic image that define the sample points, and A_θ is the affine transformation matrix dependent on θ. $T_\theta(G)$ is the grid generator, which determines the sampling of the affined image from the holistic image. Additional details, such as back propagation and sampling mechanism, can be found in [Jaderberg *et al.*, 2015]. Affined image I_a can be obtained in STN module $f_{STN}()$ as follow:

$$I_a = f_{STN}(I_{h,p}) \tag{2}$$

In this study, the localization network, the architecture of which is presented in Table 1, includes two convolutional layers (BN and ReLU layers), two pooling layers, and four fully connected layers (ReLU layers apart from the last layer). Downsampling is conducted four times by using two convolutional layers. This process is performed on the basis of two considerations: (1) the spatial information is more important than the semantic information is for the STN module, and excessive convolutional layers may lose considerable spatial information; (2) the STN module requires large receptive fields to associate the relevant patches of two images. Thus, downsampling is frequently applied. The fully connected layers estimate the 6D parameters of θ to sample the affined image from the holistic image. Such affined image is used in the ReID module.

Table 1: Architecture of our localization network. In this study, input $I_{h,p}$ is a tensor with $256 \times 128 \times 6$ shape. The designed network has few layers with a large receptive field. The FC4 layer is a regression layer and is not followed by a ReLU layer.

Name	Output Size	Parameters	Padding
Conv1	$128 \times 64 \times 16$	$[7 \times 7, 16]$, stride=2	(1,1)
Max Pool	$64 \times 32 \times 16$	2×2, stride=2	(0,0)
Conv2	$32 \times 16 \times 32$	$[3 \times 3, 32]$, stride=2	(1,1)
Max Pool	$16 \times 8 \times 32$	2×2, stride=2	(0,0)
Flatten	4096	-	-
FC1	512	4096×512	-
FC2	128	512×128	-
FC3	32	128×32	-
FC4	6	32×6	-

ReID Module

The ReID module extracts the global features of three kinds of images, namely, holistic, partial, affined images. For convenience, $f_{ReID}()$ is used to denote the ReID module. The ReID module outputs the global features and predicted person IDs in the training stage and outputs the global features only in the inference stage. In this study, ResNet50 is used as ReID module. In particular, for an arbitrary image I, we have

$$(p_I, f_I) = f_{ReID}(I) \tag{3}$$

where p_I is the predicted logits and f_I is the global feature of the image I. p_I and f_I are used to calculate the ReID loss:

$$L_R(I) = L_R(p_I, f_I) \tag{4}$$

L_R can be any ReID loss function such as L_{ID} [Zheng *et al.*, 2017] and adaptive weighted triplet loss L_{Tri} [Ristani and Tomasi, 2018] . In this study, L_R will be introduced in next section.

The ReID module has another important usage because it guides the STN module in reconstructing partial images. The ReID feature of affined image should be similar to the one of the partial image in the feature space. To achieve such target, L_{STN} is expressed as follow:

$$L_{STN}(I_p, I_a) = ||f_{I_p} - f_{I_a}||_2^2 \tag{5}$$

where f_{I_p} and f_{I_a} denote the global features of image I_p and I_a respectively. The loss of STNReID includes three ReID losses and L_{STN}, which is expressed as:

$$L = L_R(I_h) + L_R(I_p) + L_R(I_a) + L_{STN}(I_p, I_a) \tag{6}$$

3.2 Two-Stage Training

STNReID can be trained end to end in a single stage. Nevertheless, the performance of the STN module is influenced by the ReID module. An improved model with two-stage training is demonstrated in this section.

As shown as Eq.5, f_{I_p} is used to guide the STN module's training. However, such approach results in the poor performance of the STN module because the ReID module is powerful. I_p and I_a belong to the same person ID. f_{I_p} and f_{I_a} are similar in the feature space when the ReID module has a strong performance, because the features have strong clustering characteristics, especially on metric learning. In this case, the STN module cannot easily obtain knowledge. Figure **??** and Figure 4 illustrate the conclusion. Additional explanations are provided in section 5.1.

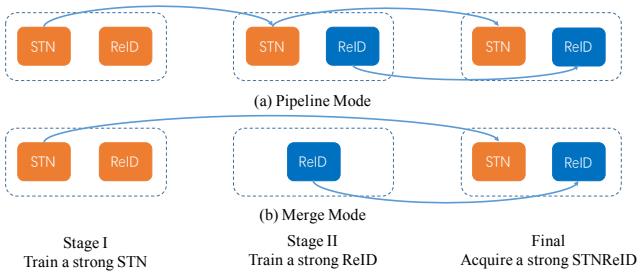


Figure 3: Two-stage training mechanism of STNReID.

As shown as Figure 3, a two-stage training mechanism is used to solve the abovementioned problem. In the first stage, STNReID is trained with a weak ReID module to acquire a strong STN module. In specifically, as shown as EP1 in Table 2, the ReID module is trained with only ID loss, and ImageNet pre-trained weights are not used to initialize it. In the second stage, the STN module of STNReID acquired from the first stage is be frozen. In addition, as shown as EP5 in Table 2, ID loss, triplet loss and some training tricks are involved to acquire a strong ReID module.

For the second stage, there are two alternative modes called Pipeline Mode (PM) and Merge Mode (MM). For the Pipeline mode (see Figure 3(a)), the ReID module in fine-tuned with the frozen STN module. Because of fine-tuning, STNReID(PM) is more suitable for the partial ReID task. However, someone may have trained a very strong ReID model on large-scale holistic or partial ReID datasets. In this case, the Merge Mode (see Figure 3(b)), which merges a STN module and a ReID module into STNReID, is free to expand a trained ReID model to a STNReID model.

The commonality between PM and MM is training a strong STN module at first, and then acquiring a strong ReID module. Finally, two-stage training mechanism let us get a strong STNReID for the partial ReID task.

4 Experimental Settings

Datasets. We use one holistic person ReID dataset and two partial person ReID datasets. Market1501[Zheng *et al.*, 2015a] is the most popular holistic dataset, and includes 32668 holistic person images of 1501 person IDs. The training set consists of 12,936 images of 751 identities. We only use the training set to train the model in our experiments. Then we test our model on two partial ReID benchmarks: Partial-ReID [Zheng *et al.*, 2015b] and Partial-iLIDs [Zheng *et al.*, 2011]. Partial-ReID includes 600 images of 60 persons, with 5 full-body images and 5 partial images per person. The images were collected at an university campus with different viewpoints, background and different types of severe occlusions (see Fig. 1). Partial-iLIDs is a simulated partial dataset based on iLIDs [Zheng *et al.*, 2011]. In the Partial-iLIDs dataset, there are 119 persons with 238 person images captured by multiple non-overlapping cameras. All the images are test data for these two partial ReID datasets. The partial

and holistic images are regarded as the probe and gallery set respectively for both datasets .

Evaluation Protocol. We follow the evaluation protocol in [He *et al.*, 2018b; Zheng *et al.*, 2015b]. We provide the average Cumulative Match Characteristic (CMC) for verification experiments to evaluate our method. Following the previous works, we train the model on Market1501 and test on partial ReID datasets.

Generate Partial Images. To train a STNReID, we should generate partials images from holistic images. Firstly, we randomly choose a direction to crop the holistic image. Then, we randomly crop $20\% \sim 60\%$ information of holistic images in the selected direction to simulate the partial images. These generated partial images have two usages, as the input partial images to train the STNReID and as augmentation to train strong baselines for partial ReID.

Settings. Our backbone network is ResNet50 [He *et al.*, 2016], which is similar to that used in most previous studies. Each image is resized into 256×128 pixels. Data augmentation involves random horizontal flipping, cropping, and generated partial images. The margin of triplet loss is 0.3, and the batch size is set to 32 with 4 randomly selected images for every 8 identities. The Adam optimizer is used, and the learning rate is 2×10^{-4} for the first 150 epochs and decays to 2×10^{-5} for next 150 epochs. Label smooth method [Szegedy *et al.*, 2016] is applied to avoid overfitting. The weight reduction is set to 5×10^{-4}. Parameters θ are initially set to $[1, 0, 0, 0, 1, 0]$.

5 Experimental Results

5.1 The ReID module affects the STN module

When end-to-end traning the STNReID, the confrontation between the STN module and the ReID module is presented, as previously discussed in Section 3.2. In this section, several experiments are conducted to verify such assumption.

To acquire ReID modules with different performances, we conduct five experiments to train the STNReID models. The training settings are shown in Table 2. However, the performance of the STN module is difficult to independently evaluate. Several aspects are considered. In particular, the rank-1 accuracy of STNReID models is assessed on the partial ReID dataset. Then, the STN modules are removed and re-evaluated, that is, only the performance of ReID modules is assessed. A performance gap is observed between such two test settings. The performance of the ReID modules from Ep1 to Ep5 strengthens gradually, and the improvements from the STN modules are smaller. However, the rank-1 accuracies should not be considered strictly because they are a normal condition in which the improvement by the algorithm becomes small given a strong baseline model. Therefore, several affine images generated by the five STN modules are visualized in Figure 4. The STN module of Ep1 accurately matches four partial images and the holistic image. The third matching pair of Ep2 slightly exhibits some redundant information. In addition, the STN module of Ep5 cannot learn much knowledge, and the affined images of the last column are similar to the holistic image.

Table 2: Training settings of five experiments. 'PT' indicates the use of pre-trained weights on ImageNet to initialize the ReID module. 'LS' denotes the label smooth trick. 'w/o STN' stands for the performance of the ReID module without the STN module. 'Imp' denotes the improvement of rank-1 accuracy from the STN module on Partial-ReID.

	PT	LS	L_{ID}	L_{Tri}	w/o STN	STNReID	Imp
Ep1	×	×	√	×	34.4	41.7	+7.3
Ep2	√	×	√	×	39.5	46.3	+6.8
Ep3	√	√	√	×	49.3	53.3	+4.0
Ep4	√	×	√	√	59.5	60.5	+1.0
Ep5	√	√	√	√	62.9	63.6	+0.8

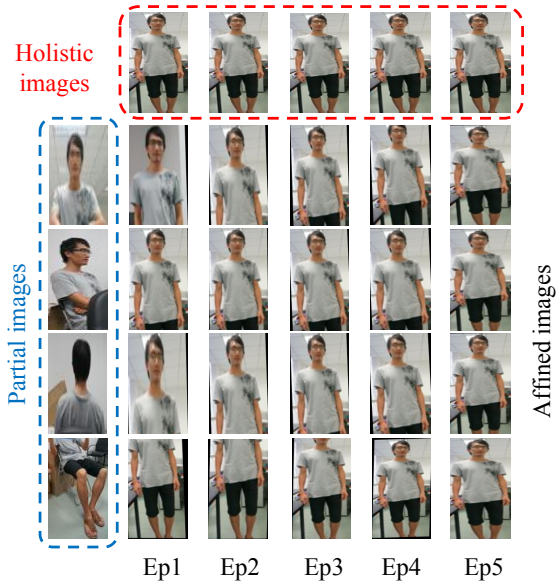


Figure 4: Sample affined images of five experiments. A holistic image and four different partial images are used. The affined images of each column belong to one certain experiment. The affined images of each row show the comparison between different experiments.

In conclusion, confrontation is observed between the STN module and the ReID modules. A strong ReID model is unsuitable for training a strong STN module in an end-to-end STNReID model with only one stage.

5.2 Ablation studies of two-stage training

In this section, STNReID in the second stage is analyzed, along with the pipeline mode (PM) and merge mode (MM). The STN module obtained from Ep1 is initialized and frozen in the second stage to train STNReID models. The results are presented in Table 3. The ReID backbone (Baseline) is implemented by the open source[1] and obtains 58.2% and 40.3% rank-1 accuracies on the partial ReID and partial iLIDS datasets, respectively.

For the MM, STNReID(MM) directly merges the ReID backbone and the frozen STN module. This mode can improve rank-1 accuracies by 3.1% and 3.4% on Partial-

[1]https://github.com/L1aoXingyu/reid_baseline

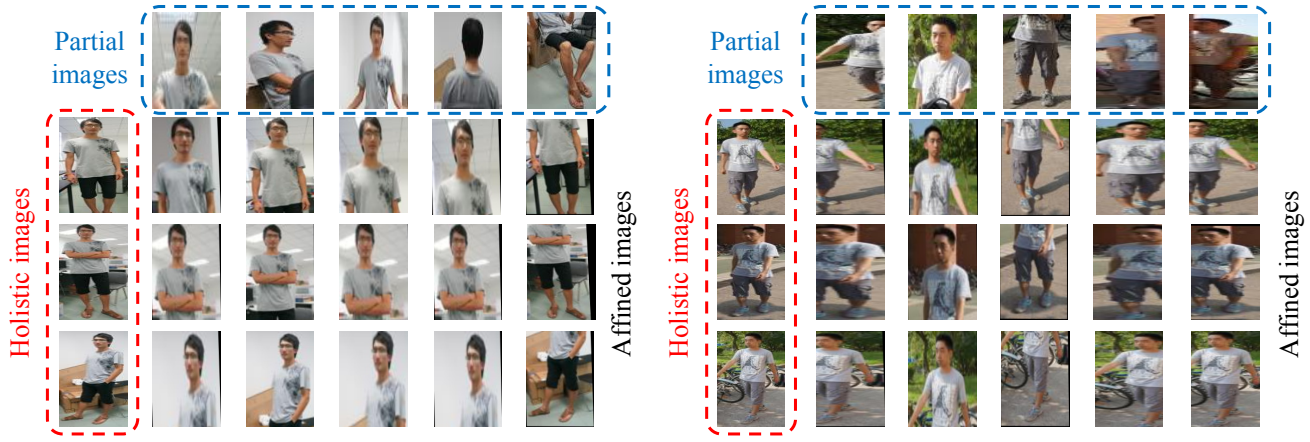


Figure 5: The example affined images of the strong STN module. We choose three holistic images and five partial images of two identities.

Model	Partial-ReID r = 1	Partial-ReID r = 5	Partial-iLIDS r = 1	Partial-iLIDS r = 5
STNReID(MM) w/o STN	58.2	82.5	40.3	71.4
STNReID(MM)	61.3	83.1	43.7	72.1
STNReID(PM) w/o STN	63.6	84.8	47.9	74.8
STNReID(PM)	66.7	86.0	54.6	78.2

Table 3: Ablation studies of two-stage training. MM and PM denote merge mode and pipeline mode respectively, and w/o STN indicates the removal of the STN module in the testing stage.

Methods	Partial-ReID r = 1	Partial-ReID r = 3	Partial-ReID r = 5	Partial-iLIDS r = 1	Partial-iLIDS r = 3	Partial-iLIDS r = 5
Resizing model	19.3	32.7	40.0	21.9	37.0	43.7
SWM	24.3	45.0	52.3	33.6	47.1	53.8
AMC	33.3	46.0	52.0	46.9	64.8	69.6
AMC+SWM	36.0	51.0	60.0	49.6	63.3	72.3
DSR(CVPR18)*	39.3	55.7	65.7	51.1	61.7	70.7
Baseline	58.2	76.5	82.5	40.3	61.3	71.4
DSR(Our)	61.7	78.9	85.3	49.6	65.3	74.8
STNReID(MM)	61.3	76.8	83.1	43.7	62.6	72.1
STNReID(PM)	66.7	80.3	86.0	54.6	71.3	79.2

Table 4: Comparison to state-of-the-art methods on Partial-ReID and Partial-iLIDS. * means the report results in original paper. However, our baseline is different from DSR's baseline. So DSR method is evaluated on our baseline for fair comparison, which is denoted as DSR(Our).

ReID and Partial-iLIDS datasets, respectively. For the PM, STNReID (PM) achieves 66.7% and 54.6% rank-1 accuracies on the partial ReID and partial iLIDS datasets, respectively; thus, it outperforms the baseline and STNReID (MM) by large margins. With the removal of the STN module, the rank-1 accuracies of non-affine features are reduced by 3.1% and 6.7% on the Partial-ReID and Partial-iLIDS, respectively. STNReID (PM) performs better than STNReID (MM) does on the partial ReID task because it fine-tunes the ReID module on partial images. Nevertheless, STNReID (MM) can be used in certain situations, such as when applications have a strong ReID model trained on a large-scale dataset, which requires considerable time and computing resources. STN module training is inexpensive.

5.3 Comparison to the State-of-the-Art

The comparison of STNReID models and state-of-the-art methods is shown in Table 4. Few studies have explored partial ReID. SWM and AMC are methods based on handcrafted features. DSR is the first deep learning-based method submitted in CVPR2018. Therefore, the comparison of our models and DSR is highlighted here. The reported rank-1 accuracies of DSR are 39.3% and 51.1% for the Partial-ReID and Partial-iLIDS datasets, respectively. Our STNReID (PM) obtains 66.7% and 54.6% rank-1 accuracies for the Partial-ReID and Partial-iLIDS, respectively. Because the training of our baseline is different from the training of DSR's baseline. DSR is evaluated with our baseline for fair comparison. The average image size of partial iLIDS is 363.3×138.3, and is thus twice or thrice larger than that of partial ReID. So that DSR method performs better on Partial-iLIDS. The performance of DSR is better than that of STNReID (MM) but is worse than that of STNReID (PM). This finding demonstrates that STNReID (PM) is a competitive method for partial ReID.

6 Conclusion, Disadvantages and Future Works

In this paper we introduced a novel deep convolutional networks with Spatial Transformer Networks (STNReID) for partial ReID. STNReID can sample the most similar patch from the holistic image to match the partial image. The experimental results demonstrate the effectiveness of our method on Partial-ReID and Partial-iLIDS. The computation costs of STNReID and DSR are also presented in Section Appendix.

However, we also thought about the disadvantages of the proposed framework. Firstly, STNReID only considers how to match the positive pairs through minimizing the L2 distances of the global features. In the inference stage, it is hard to find the accurate association between negative sample pairs. In addition, STNReID is a siamese network essentially. The siamese network is not as efficient as the one-

stream ReID model.

Although there are some shortcomings for STNReID, early experiments show the potential of it to solve the partial ReID task. We will study how to train a stronger STNReID with a simple, direct and efficient way in the future.

References

[Chen *et al.*, 2017] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proc. CVPR*, volume 2, 2017.

[Cheng *et al.*, 2016] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[Fan *et al.*, 2019] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 2019.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2018a] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7073–7082, 2018.

[He *et al.*, 2018b] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[Liu *et al.*, 2017] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017.

[Ristani and Tomasi, 2018] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046, 2018.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *The European Conference on Computer Vision (ECCV)*, September 2018.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Wei *et al.*, 2017] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 420–428. ACM, 2017.

[Xiao *et al.*, 2017] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017.

[Zhang *et al.*, 2017] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[Zheng *et al.*, 2011] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 649–656. IEEE Computer Society, 2011.

[Zheng *et al.*, 2015a] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference*, 2015.

[Zheng *et al.*, 2015b] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[Zheng *et al.*, 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[Zheng *et al.*, 2017] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Zheng *et al.*, 2018] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.