SSN3D: Self-Separated Network to Align Parts for 3D Convolution in Video Person Re-Identification

Xiaoke Jiang[1,2], **Yu Qiao** [1,3], **Junjie Yan** [2], **Qichen Li** [4], **Wanrong Zheng** [2], **Dapeng Chen** [2]

[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

[2] SenseTime Group

[3] Shanghai AI Lab, Shanghai, China

[4] MIT

{jiangxiaoke, yanjunjie, zhengwanrong, chendapeng}@sensetime.com, yu.qiao@siat.ac.cn, liqichen@mit.edu

Abstract

Temporal appearance misalignment is a crucial problem in video person re-identification. The same part of person (e.g. head or hand) appearing on different locations in video sequence weakens its discriminative ability, especially when we apply standard temporal aggregation such as 3D convolution or LSTM. To address this issue, we propose Self-Separated network (SSN) to seek out the same parts in different images. As the name implies, SSN, if trained in an unsupervised strategy, guarantees the selected parts distinct. With a few samples of labeled parts to guide SSN training, this semi-supervised trained SSN seeks out the parts that are human-understandable within a frame and stable across a video snippet. Given the distinct and stable person parts, rather than performing aggregation on features, we then apply 3D convolution across different frames for person re-identification. This SSN + 3D pipeline, dubbed SSN3D, is proved to be efficient through extensive experiments on both synthetic and real data.

Introduction

Video person re-identification (ReID) (Wang et al. 2014; Zheng et al. 2016; Hou et al. 2019) is of great interests as it provides the temporal variant appearance of a person to achieve more accurate ReID. In this task, a query video clip is given to find the clips belonging to the same person from a large video gallery. In order to aggregate useful information, LSTM (Hochreiter and Schmidhuber 1997; Yue-Hei Ng et al. 2015) and 3D convolutional networks (Tran et al. 2015; Carreira and Zisserman 2017; Qiu, Yao, and Mei 2017) are widely employed. And 3D convolution is believed to outperform LSTM on video person classification (Carreira and Zisserman 2017). However, 3D convolution suffers from temporal misalignment issue. It processes the features at the same spatial position in adjacent frames into one value, which expects the human pose and camera viewpoints to be aligned before being fed into the network. When we observe many video person ReID datasets, it is manifest that due to the imperfect person detection algorithm, the locations where a person appears in the bounding boxes are inconsistent, not to mention the pose variation of person.

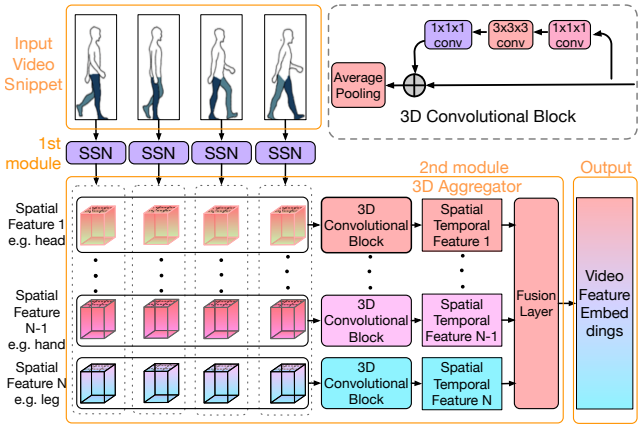

Figure 1: Overall Workflow of SSN3D. It takes a video snippet as input, and outputs a feature embedding vector

In this paper, we address the appearance misalignment for 3D convolution. The whole network includes two modules as shown in figure 1. One named self-separated network (SSN), leveraging attention mechanism, seeks out distinct person parts while preserving the original spatial information within a single image. SSN is then applied to a video snippet, and same parts appeared in multiple frames are aligned forming 4D-tensors. The other is a 3D convolution based aggregator. It takes the 4D-tensors of each part as input, summarizes the feature of distinct parts through the temporal dimension, and transforms the resulted feature maps into a single feature vector. To indicate the combination of SSN and 3D convolution, we name our method **SSN3D**.

SSN is different from conventional attention mechanism. A sophisticated two-round classification is designed to take advantage of the consistency between the pixel-wise feature and the aggregated feature. The consistency provides signals for unsupervised training, and allows SSN to separate those distinct parts of person with cross-entropy loss. That's why we call it self-separated network. However, unsupervised training, without human's guidance, would find some parts that are distinct, but cannot guarantee those parts to be human-understandable or stable due to its blind searching.

Alternatively, we enable supervised training, which relies on person's key points as anchor, making the selected parts more human-understandable. However, employing an existing pose detection model (He et al. 2017; Cao et al. 2018) is time-consuming and lacks robustness due to the imperfection of the model. To combine the advantages of the above two strategies, we use some high-quality labeled data to guide SSN training, and also feed the network with unlabeled data. The selected parts in this semi-supervised strategy are not only distinct and human-understandable within a frame, but also stable across different frames. The combined strategy outperforms both the unsupervised strategy and the supervised strategy(labeled via OpenPose (Cao et al. 2018)) on ReID task.

SSN+3D is also a novel combination for temporal aggregation. Instead of performing separating and summarizing within one image, SSN+3D essentially involves the 3D convolution into each attention channel. Unlike the popular design of fusing different parts within a frame via pooling (Liu et al. 2017; Zheng et al. 2016), we align different parts, and construct 4D-tensors for each part across multiple frames. For each 4D-tensor, a corresponding learnable 3D convolutional block is used to aggregate the feature of the part it represents. Feature fusion is conducted after the features of all distinct parts are extracted. Compared to existing methods, SSN+3D implements finer-grained part alignment and feature aggregation, allowing to detect variation of parts in both spatial and temporal dimensions.

To sum up, the contributions of this paper are three-fold: (1) We propose SSN+3D pipeline to address the appearance misalignment problem in video person ReID. SSN+3D provides extraordinary ability to handle the both spatial and temporal variation of person parts. Our final solution can achieve superior performance to state-of-the-art methods on iLIDS-VID and DukeMTMC, and comparable performance on MARS. (2) SSN together with two-round classification mechanism, seeks the consistency between the pixel-wise feature and the aggregated feature in different training strategies. This design can be generalized to train different attention networks. (3) Compared to existing pooling-based aggregator and 3D convolution on fused feature, our 3D convolution based feature aggregator is finer-grained, and adapts to the variation of person parts better.

Related Work

Video Person ReID: The samples of a Video ReID task contain more frames and additional temporal information compared with image ReID. Usually, for each individual, there are multiple video clips, and we compare each piece of video against each other to determine whether they belong to the same person. Multiple models have been proposed to consider the temporal information. (Wang et al. 2014) chooses the frames with maximum and minimum flow energy, while (McLaughlin, Martinez del Rincon, and Miller 2016) and (Yan et al. 2016) use RNN to make use of such information. Additionally, (Li, Zhang, and Huang 2019) proposes a two-stream convolution network to extract spatial and temporal cues for video person ReID, and (Chung, Tahboub, and Delp 2017) propose a two-stream Siamese CNN which processes spatial and temporal information separately. Besides, images ReID models (Liu, Yan, and Ouyang 2017; Li et al. 2018; Si et al. 2018; Chen et al. 2018), when integrated with multi-frame features, can still be very successful on video task. In this paper, we leverage attention mechanism to align each parts of person, which, when working with a 3D convolutional networks, can deeply benefit the quality of the feature extracted from the video clips.

Recent Progress: MG-RAFA(Zhang et al. 2020) trains pixel-wise weights to extract attention map from the original feature map via supervised learning strategy. Average pooling is used to build the original reference of feature to construct attention map. Both MGH(Yan et al. 2020) and ST-GCN(Yang et al. 2020) leverage graph convolution approach to model the relationship between the parts of intra-/inter- frames in a video snippet. Both leverage PCB-like mechanism to split feature map to multiple parts, which then play the role of nodes of GCN. ST-GCN models the relationship directly via building multiple graphs. MGH creates PCB-like feature parts in different granularity, e.g. global feature, local feature with two/four partitions, and then builds hyper-graph to explore the spatial and temporal relationships.

AP3D(Gu et al. 2020) and SSN3D both have body alignment module and feature aggregation module, respectively. But they are very different. For the body alignment module, AP3D provides a person feature by cross-pixel semantic similarity, while SSN3D align different person parts with cross-entropy loss and two-round classification.

For the feature aggregation module, AP3D only has one branch to capture the temporal variation of the global features from different frames. While SSN3D adopts different 3D CNN to aggregate temporal feature.

Both SSN3D and SpaAtn(Li et al. 2018) have an attention mechanism but in different ways. i.e. the cross-entropy loss and two-round classification, instead of KL-divergence as SpaAtn. Furthermore, our performance is better than SpaAtn by a large margin on MARS and iLIDS-VID. (SpaAtn does not provides results on DukeMTMC-VideoReID). The essential reason is that KL-divergence cannot tell the differences between similar parts, while the two-round classification under semi-supervised training strategy can.

SSN aims to align person parts wherein key points of the pose are used as anchors, and the probability of each pixel belonging to different anchors is inferred though attention classifier (unsupervised/semi-supervised trained). Even though attention mechanism, pose detection and 3D CNN are not firstly utilized in the Video ReID field, using the combination of cross-entropy and two-round classification to realize attention map is new to this field. We combine the existing ideas to present a simple yet effective method and experiment results demonstrate the competing performance.

The Proposed Approach

Overall framework of our feature extraction model is presented in figure 1, which is composed of two modules, SSN and 3D convolutional feature aggregator. SSN seeks out N distinct parts of person from each image. The same category of parts (e.g. head) from a video snippet are aligned forming

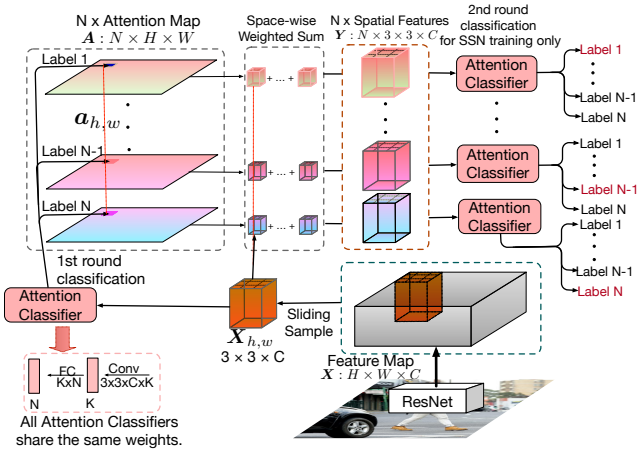

Figure 2: Structure of SSN. It takes an image as an input, and outputs N spatial features. Each pixel of X is estimated twice by the attention classifier

a 4D-tensor. The N 4D-tensors are then fed to N 3D convolutional blocks for feature aggregation. The entire system generates the video embedding in an end-to-end fashion.

Self-Separated Network

SSN, shown in figure 2, takes an image as an input and outputs N spatial features of size $3 \times 3 \times C$ (in our case, $N = 9, C = 1024$). We use ResNet50(He et al. 2016) as the backbone of the convolutional layers, whose output is feature of size $H \times W \times C$. The features from the sliding window is fed into an attention classifier (1st round classification), which is composed of a convolutional layer followed by a fully connected layer that can map the features to a lower dimension. After a softmax operation on the attention maps, the sum of all elements in each attention map will be 1. Each component of this attention map describes the weight of its corresponding sliding window in the feature map. After calculating the weighted sum of the sliding window features based on, we can output N spatial features.

For SSN training, the produced spatial features will be fed into the same attention classifier with a softmax layer (2nd round classification). The result of 2nd round classification, which creates labels for unsupervised training, should be consistent with the 1st round classification. We then use cross-entropy loss to guarantee that each label describes different and distinct part, even if we do not know what these parts represent. Our experiments show that the loss will converge on a small value, indicating that the model does learn the distinct parts. Moreover, we can also use an semi-supervised learning method to instruct the targets to be the parts we desire. In the following subsections, we will explain the figure 2 in detail.

Network Architecture. We introduce the components of SSN as follows.

Attention Classifier. The attention classifier demonstrated in figure 2 maps the convolution features to a N dimensional vector. Our attention classifier has a convolutional layer, which uses 3×3 kernel with K output channels. This convolutional layer maps the features to a vector with K elements (in our case, $K = 512$). After a relu layer, the vector is then fed into a $K \times N$ forward function. We define

$$\boldsymbol{a}_{h,w} = AttentionClassifier(\boldsymbol{X}_{h,w}), \tag{1}$$

where $\boldsymbol{a}_{h,w} \in \mathbb{R}^N$ is the output, and $\boldsymbol{X}_{h,w} \in \mathbb{R}^{3\times3\times C}$ is the feature of sliding sample from the feature map at the position (h, w).

In the 1st round classification, we apply the attention classifier to the pixel of input feature. Softmax is not applied to this result, which implies the original possibilities of pixel belonging to different labels are kept this turn.

Attention Map. We have N attention maps generated from the feature map, where the i-th is computed by $A^i \in \mathbb{R}^{H\times W}$. Each attention map is computed by the previous attention classifier, i.e. $A^i_{h,w} = \boldsymbol{a}^i_{h,w}$, where $\boldsymbol{a}^i_{h,w}$ is the i-th element of $\boldsymbol{a}_{h,w}$.

In the previous section, we have explained that the attention classifier has not been applied with a softmax layer in the 1st round classification, and we do it when attention map is used to calculate the corresponding weights with the sliding window features. A element of final output from the attention map $\hat{A}^i$ is calculated as

$$\hat{A}^i_{h,w} = \frac{\exp(A^i_{h,w})}{\sum_{h=1}^{H}\sum_{w=1}^{W}\exp(A^i_{h,w})}. \tag{2}$$

Space-wise Weighted Sum. We now have N attention maps for N output features. As is mentioned before, each attention map provides a set of weights for the sliding window features. We use

$$\boldsymbol{Y}^i = \sum_{h=1}^{H}\sum_{w=1}^{W}\hat{A}^i_{h,w} \cdot \boldsymbol{X}_{h,w} \tag{3}$$

to produce the i-th of the final N tensor features for unsupervised learning.

Training with 2nd Round Classification. Now that we have N spatial features, and then the next concern is how to train our model. Here we apply the identical attention classifier to the N spatial feature, and result of 2nd round classification should be consistent with the 1st. For example, the spatial feature generated via the attention map of label 1, after another attention classifier, should still belong to the label 1. Cross-entropy loss is used to train the model. The nature of the cross-entropy loss guarantees that each label describes different information, because when training, it will maximize one and suppress the rest. As a result, the spatial features across the frames belonging to the same label express homogeneous information, which, in fact, is the alignment operation we demonstrated before. Generally, we will set the target of the first feature output to be label 1, the target of the second to be 2, and so on. If the loss converges to a small value, it means the model has indeed learned something. So we define the loss as

$$\mathcal{L}_{SSN} = -\sum_{i=1}^{N}\log\left(\frac{\exp(p_{i,i})}{\sum_{j=1}^{N}\exp(p_{i,j})}\right), \tag{4}$$

where $p_{i,j}$ is the predicted probability of part i is of category j. Note that the 2nd round classification is used for training only.

With two-round classification and cross-entropy loss SSN can be trained in the unsupervised strategy. But this unsupervised training is more or less "blind". The attention classifier is instructed by a consistent signal without knowing what it is, be it a head or hand. Essentially, it is to leverage the similarity of pixel values (RGB feature or high-dimension deep feature). However, without giving a specified part to select, parts with similar pixel values may interfere with each other. As a consequence, unsupervised learning seeks out the parts that are distinct but may locate in different positions if they are similar with each other. Using pixel similarity is widely accepted in the context of ReID(Shen et al. 2018; He et al. 2019), since pixel value is one of the most discriminative feature. This two-round classification unsupervised training scheme has been proved by our later experiment to be so powerful that it could even select distinct pattern from the samples generated by Gaussian distribution, but not good at handling similar person parts. So we revise our method to enable semi-supervised training for better performance.

Semi-supervised Training. In the above text, by two-round classification, we are able to train SSN in an unsupervised strategy. However, we do not know what each spatial feature represents, whether label 1 or label 2 is the person's hand or head. Here we proposal semi-supervised strategy to instruct the targets to be the parts we desire. The idea is very simple, we feed the network with labeled and unlabeled data at the same time. labeled data plays the role of "anchor", which is a strong signal to instructs attention classifier to select those N parts that we want. In details, we mix some features bypassing our SSN into those features produced by SSN. That is to replace Y^i in equation 3 with $\boldsymbol{Y}^i = \boldsymbol{X}_{h_i,w_i}$, where (h_i, w_i) is our manually labeled key point of the i-th target. Therefore, labeled data and unlabeled data can be fed to the network together.

In this way, semi-supervised SSN seeks out person parts which is distinct within a frame and stable across multiple frames.

3D Convolutional Network Block

As we have aligned the spatial features by our SSN, the 3D convolutional blocks can now effectively extract temporal information. Figure 1 has demonstrated that after each frame has been fed into a SSN, N spatial features in the size of $3 \times 3 \times C$ have been produced. We then concatenate those spatial features belonging to the same label in the temporal dimension to produce a feature (4D-tensor) in the shape of $T \times 3 \times 3 \times C$ (T is the number of pictures in the snippet. In our case, $T = 4$). Those features would be fed into a 3D convolutional block for feature extracting. Note that the 3D convolutional blocks do not share weights since they are distinct parts. After all N sets of spatial features have been mapped to a one-dimensional vector by the 3D convolutional networks, we concatenate all of them and apply a fusion forward network to produce the final embedding feature for the video.

Training with Hard Samples Mining. Unlike the common practice used by (Wang et al. 2018), who combines classification loss and triplet loss for spatial-temporal representation learning. We discard the classification loss which considering person identities as category labels and replace it with the cross-entropy loss from our SSN.

Triplet loss with hard samples mining (Hermans, Beyer, and Leibe 2017) is a common practice for person ReID task, and it is described as

$$\mathcal{L}_{tri} = \sum_{i=1}^{B} [m + \max_{f_p \in S_i^+} \frac{||f_i - f_p||_2}{\sqrt{d}} - \min_{f_n \in S_i^-} \frac{||f_i - f_n||_2}{\sqrt{d}}]_+, \tag{5}$$

where m is a pre-defined margin, d is the dimension of the output features, f_i is the video feature of the i-th sample, and $[\cdot]_+ = max(0, \cdot)$. S_i^+ and S_i^- are the positive and negative sample sets of the i-th sample respectively.

Loss Function

The final objective function $\mathcal{L}$ is formulated as the weighted sum of the SSN loss and the triplet loss, i.e.

$$\mathcal{L} = \mathcal{L}_{tri} + \lambda \cdot \mathcal{L}_{SSN}. \tag{6}$$

The reason why we give our SSN loss a coefficient λ is that it converges very fast, which may impede our learning.

Experiment

In this section, we will first evaluate SSN+3D on video person ReID tasks, then evaluate our core module, SSN, on both synthetic and real data.

Experimental Setting

Datasets. We evaluate our method in three video person ReID datasets. In particular, *iLIDS-VID* consists of 600 video sequences, with 300 different individuals captured by two cameras. The length of the video sequence varies from 23 to 192 frames. *MARS* has 1,261 identities with more than 20,000 video sequences captured from 6 cameras. Bounding boxes are produced by DPM detector (Felzenszwalb et al. 2009) and GMMCP tracker (Dehghan, Modiri Assari, and Shah 2015). *DukeMTMC-VideoReID* is a subset of the tracking DukeMTMC (Ristani et al. 2016) benchmark, and we use DukeMTMC as its name for short. The pedestrian images are cropped from the video for 12 frames every second to generate a tracklet.

Evaluation Protocols. We employ the Mean Average Precision (mAP) (Zheng et al. 2015) and Cumulative Matching Characteristics (Bolle et al. 2005) for evaluation.

Implementation Details. *CNN Backbones:* We use the pre-trained ResNet50 for the convolution layers. Note that each ResNet model has a total of 5 layers, we only use the first four of them, which produce feature maps of 1024 channels. *Supervised Labels:* For the supervised labels in each dataset, we use OpenPose(Cao et al. 2018) to mark the position of their head, body, crotch, left and right elbows, knees and feet. When being trained, if a snippet is selected, their corresponding labeled images, along with their labels, will also be

Learning Strategy	iLIDS		MARS		DukeMTMC	
	top-1	mAP	top-1	mAP	top-1	mAP
Supervise	73.4	75.8	69.8	61.1	86.3	79.6
UnSuperv.	83.1	84.2	82.4	67.5	89.9	86.2
Semi-Sup.	**88.9**	**89.2**	**90.1**	**86.2**	**96.8**	**96.3**

Table 1: Different learning strategy employed by SSN. It is clear to see that semi-supervised learning is superior to others in all the person ReID tasks

Attention Classifiers	iLIDS		MARS		DukeMTMC	
	top-1	mAP	top-1	mAP	top-1	mAP
NonSha.	69.4	73.2	72.3	70.9	84.9	71.2
Sharing	**88.9**	**89.2**	**90.1**	**86.2**	**96.8**	**96.3**

Table 2: Comparison of with and without weight sharing

selected. *Final feature representation:* The architecture of the 3D CNN is demonstrated in figure 1. The fusion layer casts the nine concatenated features to a 1024-dimension vector. *Training and testing protocols:* In the training stage, for each video tracklet, we randomly sample 4 frames with a stride of 8 frames to form a video clip. Each batch contains eight individuals, and each individual has four video snippets. All the input images are resized to 256×128 pixels. Adam (Kingma and Ba 2014) with a weight decay of 0.0005 is adopted to update parameters. The network is trained for 150 epochs in total, with an initial learning rate of 3×10^{-4}. Learning rate is reduced with a decay rate of 0.1 after 50 epochs. In the testing stage, each video tracklet is split into multiple 32-frame video clips. Then we extract the feature representation for each video snippet and use their average to represent them.

Ablation Study on SSN+3D Pipeline

Different Learning Strategy. The supervised learning is that, instead of using attention maps to calculate the spatial features of each frame, we directly feed labeled data into following 3D Convolutional Networks for feature extracting. The unsupervised training, however, is to let the model find the distinct parts itself without any labels. The semi-supervised learning is that we select two images with high-quality labels from each video tracklet. And the training method is described in the previous section.

As is demonstrated in table 1, the best way of training our model is to provide some high-quality labels and let the model learn the rest. This way of semi-supervise learning instructs our attention classifiers to pay attention to our desired body parts. Because of the nature of the corrupted data, the OpenPose model may fail on some of the images, which we suppose is the reason why the supervised learning does not work as good as the others. The performance of unsupervised learning is not bad, but it still lags behind semi-supervised learning due to lack of guidance. We set $\lambda = 0.1$ on iLIDS and MARS, and $\lambda = 0.05$ on DukeMTMC.

If without explicit declaration, the following experiments use the semi-supervised strategy.

Weight Sharing. In our design, the spatial attention classifiers for a single frame and the temporal attention classifiers

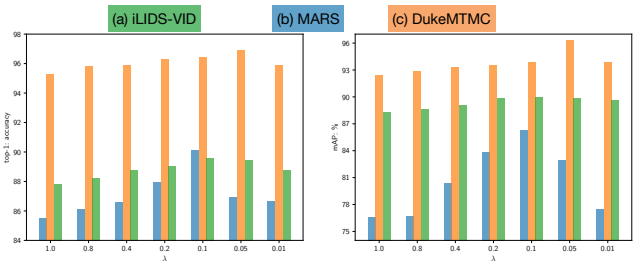

Figure 3: The top-1 and mAP on (a)iLIDS-VID, (b)MARS(b), and (c)DukeMTMC

Methods	top-1	top-5	top-10
LFDA	32.9	68.5	82.2
KISSME	36.5	67.8	78.8
LADF	39.0	76.8	89.0
STF3D	44.3	71.7	83.7
TDL	56.3	87.6	95.6
MARS	53.0	81.4	-
SeeForest	55.2	86.5	91.0
CNN+RNN	58.0	84.0	91.0
Seq-Decision	60.2	84.7	91.7
ASTPN	62.0	86.0	94.0
QAN	68.0	86.8	95.4
RQEN	77.1	93.2	97.7
STAN	80.2	-	-
Snippet	79.8	91.8	-
Snippet+OF	85.4	96.7	**98.8**
VRSTC	83.4	95.5	97.7
AP3D	86.7	-	-
SSN3D	**88.9**	**97.3**	**98.8**

Table 3: Comparison with related methods on iLIDS-VID

across multiple frames share the same weight. However, we can loose this constraint and explore what outcomes it will bring to us. When not sharing attention weights, each spatial feature is a certain combination of different parts of the image in a way we cannot control, neither do we know what each spatial feature really represents. Table 2 has proved to us the benefit of this weight sharing between the spatial and temporal classifier. When not sharing weights, T independent SSN losses are used during training. λ is set to 0.1 on iLIDS and MARS, and to 0.05 on DukeMTMC when running experiments. The large margin between with and without weight sharing implies the impact of SSN.

Influence of the Parameters λ. λ is the parameter to balance the relative effects of the SSN loss. We analyze the impact of the λ on iLIDS-VID, MARS, and DukeMTMC datasets respectively. We observe that our method achieves the best performance when we set λ to be relatively small, e.g. 0.1 or 0.05. This is because the SSN's strong learning ability may lead to a fast convergence, which is harmful when not enough data has been represented. We demonstrate our results in figure 3.

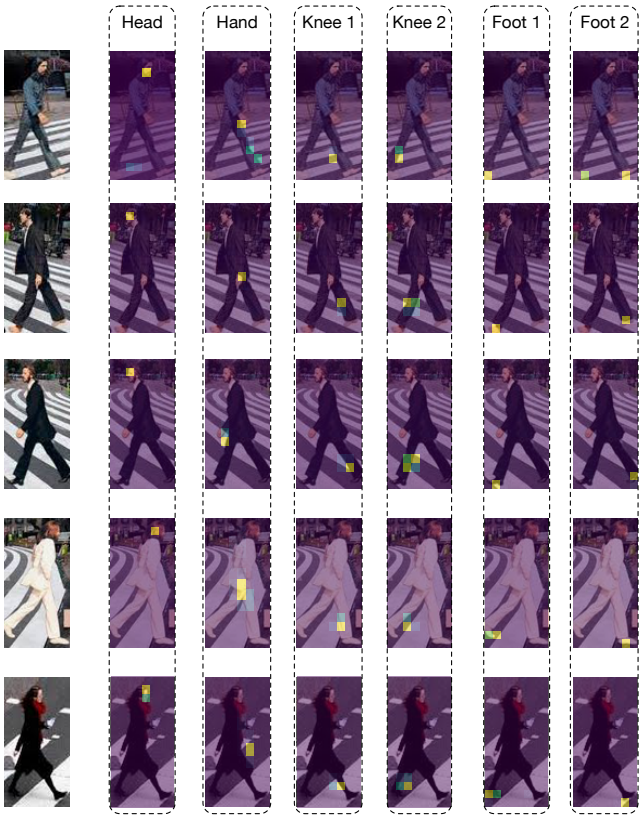

Figure 4: Finding Specified Parts in Semi-supervised Strategy. The above image demonstrates the result of attention maps on the Pedestrian 128 dataset. The first pedestrian is labeled, and the rest is unlabeled. It is clear to see that each attention map has successfully learned where to focus on each individual

Comparison with State-of-the-arts

Table 3, 4, and 5 report the performance of our approach and other state-of-the-art methods on iLIDS-VID, MARS, and DukeMTMC benchmark respectively. The methods in these tables differ in many aspects, from deep models to traditional models. On iLIDS-VID, SSN3D outperforms others in terms of top-1, top-5, and top-10 CMC. On DukeMTMC, SSN3D outperforms others in top-1, top-10 and mAP. And on MARS, SSN3D also achieves comparable scores with state-of-the-art methods. We believe this improvement comes mainly from our SSN+3D pipeline. Impressive they are, we do not regard these scores as the only judgement to the model we propose. And SSN and two-round classification is what we value most. Therefore, we are expecting to more practical applications of our SSN to other tasks besides video person ReID.

Study on SSN

In this subsection, we study our core design, SSN. To prove the convergence ability, and demonstrate the results of our model trained by unsupervised, and semi-supervised learning, we prepare the following three datasets for each task. Considering the different nature of the datasets, we use different model settings to secure visualized results. We use the same

Methods	top-1	top-5	top-10	mAP
Mars	68.3	82.6	89.4	49.3
SeeForest	70.6	90.0	97.6	50.7
Seq-Decision	71.2	85.7	91.8	-
Latent Parts	71.8	86.6	93.0	56.1
QAN	73.7	84.9	91.6	51.7
K-reciprocal	73.9	-	-	68.5
RQEN	77.8	88.8	94.3	71.7
TriNet	79.8	91.3	-	67.7
EUG	80.8	92.1	96.1	67.4
STAN	82.3	-	-	65.8
Snippet	81.2	92.1	-	69.4
Snippet+OF	86.3	94.7	**98.2**	76.1
VRSTC	88.5	96.5	97.4	82.3
AP3D	**90.1**	-	-	85.1
SSN3D	**90.1**	**96.6**	98.0	**86.2**

Table 4: Comparison with related methods on MARS

Methods	top-1	top-5	top-10	mAP
EUG	83.6	94.6	97.6	78.3
VRSTC	95.0	**99.1**	**99.4**	93.5
AP3D	96.3	-	-	95.6
SSN3D	**96.8**	98.6	**99.4**	**96.3**

Table 5: Comparison with related methods on DukeMTMC

Adam optimizer (Kingma and Ba 2014) with weight decay 5×10^{-4} to update the parameters. Note that we evaluate SSN only without 3D part.

Random Generation. To prove the strong learning ability of our models, we use data generated by different random distribution, including Gaussian, and Uniform distribution. The random distribution generates a random sample in the size of $128 \times 16 \times 16$, and our sliding window size is set to be 3×3. We do not use any convolutional backbones for feature extraction, and we feed the sample directly into the attention classifier. We train the model with the batch size of 128.

As is demonstrated in figure 5, we run a total number of 9 experiments for this dataset. It is crystal clear that the learning curve of our model running on random data has convinced us of the substantial learning ability of our model. The loss can converge to a small value very fast even the data is randomly generated, especially when the number of labels N is small and the data is more dynamic (data generated from a normal distribution is much dynamic than that generated from a 0-1 uniform distribution). It is reasonable that the learning curves of the all-one feature maps do not converge because it is impossible to spot different patterns from homogeneous information. Another thing worth noticing is that it seems harder for our model to learn when the number of labels increases. We suppose that this is because it is harder to spot more different parts in a feature map with a fixed size. All in all, this experiment on random data has revealed the extraordinary learning ability of our model.

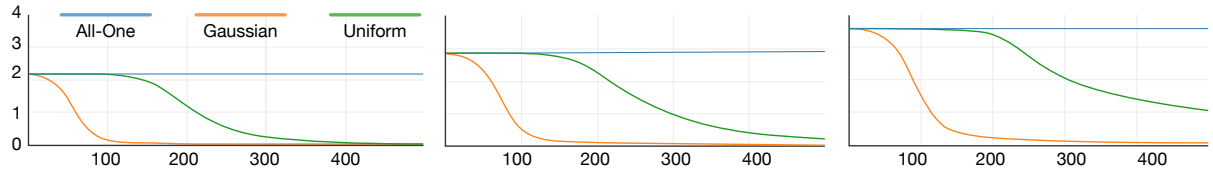

Figure 5: Strong Learning Ability. The x-axis is the training iterations, and the y-axis is a total loss

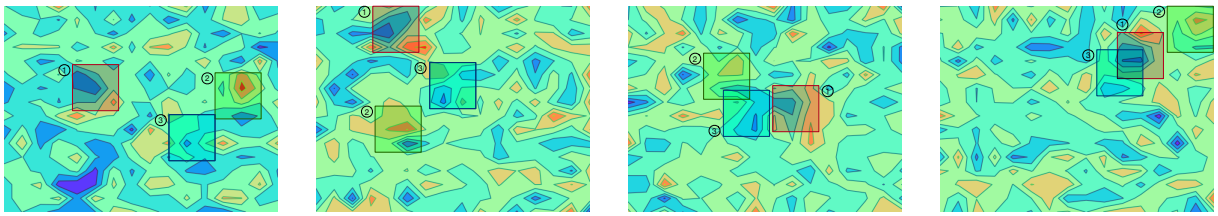

Figure 6: Performance of Unsupervised Learning. Our SSN model is able to spot similar parts between different paintings when appreciating our Amber Abstract dataset. Type 1 captures the pattern with a light blue protuberance on the left; type 2 spots the feature of an orange diamond above a blue stripe; and each type 3 contains a small deep blue block

Amber Abstract. To explore the performance of our model trained by unsupervised learning, we use this dataset for the part alignment task. We produce this dataset in digital abstract expressionism since there are not many ready-made datasets available for such a challenge. Inspired by the shapes and colors of amber, we paint 128 different paintings in the same style. While creating these works, we do not have any intentions to draw similar parts among different pieces, therefore there is no "right answer" for the shared areas. We use this dataset to give readers a visualized and direct sense of the effectiveness of our SSN trained by unsupervised learning. We call this dataset Amber Abstract. We resize the original painting in the size of 480×640 to the size of 120×160, we did not use any convolutional backbones for feature extraction. We feed the painting directly into the attention classifier, which maps each $3 \times 19 \times 19$ sliding window to a 1×9 vector. The batch size is 4.

Some of our results on Amber Abstract dataset are shown in figure 6. To provide readers with a more direct sense of what our model produces, we let our model find aligned parts, which are similar parts from different random creations. There is no labeled data in the training process, and our model can spot these parts, as is demonstrated in bounding boxes of different colours, without efforts. This whole learning process is unsupervised, and we will display more of the results in the appendix.

Pedestrian 128. The semi-supervised learning results are demonstrated through the 128 pedestrian bounding boxes in the shape of 60×120 we collected from the Internet. We manually labeled 16 of them (not all of them) with their body key points including head, hands, knees and feet. We use 3 blocks of ResNet34 to extract a $128 \times 8 \times 15$ feature map for each image. The sliding window is in the size of 3×3.

During training, we use the labeled images to guide SSN training. After that, we continue training SSN with unlabeled data. We demonstrate some of the attention maps of different labels, along with their original images, in figure 4.

We also train SSN in unsupervised strategy. As we have expected, the unsupervised SSN selects distinct parts, but some parts are not human-understandable. Also, the selected parts is not as stable or accurate as in the semi-supervised strategy.

SSN Study Summary In this part, we take SSN out from the pipeline and verify its robustness through three datasets. We show its strong learning ability of seeking out distinct parts, its flaws when trained in unsupervised strategy, and a better result in semi-supervised strategy.

Conclusion

In this paper, we propose a novel design, called SSN3D. SSN and two-round classification provide a general way of training attention networks in different strategies (and of course with different performances). The main advantage of SSN is to guarantee the distinction and stableness of spatial and temporal information. The 3D convolution based aggregator shows an extraordinary capacity to handle the temporal variation of different parts.

In practical applications, this whole system achieves impressive results on video person ReID tasks. As future work, we are expecting to see more forms of such an alignment model on other tasks. Also, we plan to explore more flexible semi-supervised learning strategy, e.g. with less or partial labeled data.

Acknowledgements

Project supported by the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800).

References

Bolle, R. M.; Connell, J. H.; Pankanti, S.; Ratha, N. K.; and Senior, A. W. 2005. The relation between the ROC curve and the CMC. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, 15–20. IEEE.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* .

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1169–1178.

Chung, D.; Tahboub, K.; and Delp, E. J. 2017. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1983–1991.

Dehghan, A.; Modiri Assari, S.; and Shah, M. 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4091–4099.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32(9): 1627–1645.

Gu, X.; Chang, H.; Ma, B.; Zhang, H.; and Chen, X. 2020. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In *ECCV*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

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

He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; and Feng, J. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 8450–8459.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* .

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7183–7192.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8618–8625.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 369–378.

Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; and Feng, J. 2017. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology* 28(10): 2788–2802.

Liu, Y.; Yan, J.; and Ouyang, W. 2017. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5790–5799.

McLaughlin, N.; Martinez del Rincon, J.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1325–1334.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, 5533–5541.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, 17–35. Springer.

Shen, Y.; Xiao, T.; Li, H.; Yi, S.; and Wang, X. 2018. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6886–6895.

Si, J.; Zhang, H.; Li, C.-G.; Kuen, J.; Kong, X.; Kot, A. C.; and Wang, G. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5363–5372.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *European conference on computer vision*, 688–703. Springer.

Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; and Yang, X. 2016. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, 701–716. Springer.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2899–2908.

Yang, J.; Zheng, W.-S.; Yang, Q.; Chen, Y.-C.; and Tian, Q. 2020. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3289–3299.

Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4694–4702.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2020. Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10407–10416.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 868–884. Springer.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.