VRSTC: Occlusion-Free Video Person Re-Identification

Ruibing Hou[1,2], Bingpeng Ma[2], Hong Chang[1,2], Xinqian Gu[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
{ruibing.hou, xinqian.gu}@vipl.ict.ac.cn, bpma@ucas.ac.cn, {changhong, sgshan,xlchen}@ict.ac.cn

Abstract

Video person re-identification (re-ID) plays an important role in surveillance video analysis. However, the performance of video re-ID degenerates severely under partial occlusion. In this paper, we propose a novel network, called Spatio-Temporal Completion network (STCnet), to explicitly handle partial occlusion problem. Different from most previous works that discard the occluded frames, STCnet can recover the appearance of the occluded parts. For one thing, the spatial structure of a pedestrian frame can be used to predict the occluded body parts from the unoccluded body parts of this frame. For another, the temporal patterns of pedestrian sequence provide important clues to generate the contents of occluded parts. With the spatio-temporal information, STCnet can recover the appearance for the occluded parts, which could be leveraged with those unoccluded parts for more accurate video re-ID. By combining a re-ID network with STCnet, a video re-ID framework robust to partial occlusion (VRSTC) is proposed. Experiments on three challenging video re-ID databases demonstrate that the proposed approach outperforms the state-of-the-arts.

1. Introduction

Video person re-identification (re-ID) aims at matching the same person across multiple non-overlapping cameras, which has gained increasing attention in recent years. However, it remains a very challenging problem due to large variations of appearance caused by camera viewpoints, background clutter, and especially partial occlusion. The performance of video re-ID usually degenerates severely under partial occlusion. This problem is difficult to tackle as any part of the person may be occluded by other pedestrians and environmental objects (*e.g.* bicycles and indicators).

Typical video re-ID methods [21, 30, 32] do not take into account the effect of partial occlusion. They represent each frame of a video as a feature vector and compute an aggregate representation across time with average or maximum pooling. In the presence of partial occlusion, the video feature is usually corrupted due to the equal treatment of all frames, leading to severe performance degeneration.

Recently, the attention mechanism has been introduced to video re-ID in order to deal with partial occlusion [18, 43, 33, 15, 3]. They select discriminative frames from video sequences and generate informative video representation. Although these approaches have a certain tolerance to partial occlusion, it is not ideal to discard the occluded frames. On one hand, the remaining visible portions of the discarded frames may provide strong cues for re-ID. So these methods lost too much appearance information in video features, making them difficult to identify the person. On the other hand, the discarded frames interrupt the temporal information of video. The works [21, 30, 32] have verified that the temporal information of video can help to identify the person. For instance, if different persons have similar appearance, we can disambiguate them from their gaits. Therefore, these methods may still fail when the partial occlusion occurs.

In this work, we propose Spatial-Temporal Completion network (STCnet) to explicitly tackle the partial occlusion problem by recovering the appearance of the occluded parts. For one thing, according to the spatial structure of pedestrian frame, the visible (unoccluded) body parts can be used to predict the missing (occluded) body part of a person. For another, because of the temporal patterns of pedestrian sequence, the information from adjacent frames is helpful for recovering the appearance of the current frame. Motivated by the two facts, we design the spatial structure generator and temporal attention generator in STCnet. The spatial structure generator exploits the spatial information of the frame to predict the appearance of the occluded parts. The temporal attention generator exploits the temporal information of the video with a novel temporal attention layer to refine the parts generated by the spatial generator. With the spatial and temporal generators, STCnet is able to recover the occluded parts.

Furthermore, we propose an occlusion-free video re-ID framework by combining a re-ID network with STCnet (VRSTC), where the unoccluded frames are used to train and test the re-ID network. Due to the superior completion ability of STCnet, the video re-ID framework, VRSTC, achieves robustness to partial occlusion. We demonstrate the effectiveness of the proposed framework on three challenging video re-ID datasets, and our method outperforms the state-of-art methods under multiple evaluation metrics.

2. Related Works

Person re-identification. Person re-ID for still images has been extensively studied [41, 20, 16, 37, 14, 38, 28]. Recently, researchers start to pay attention to video re-ID [17, 21, 30, 32, 43, 15, 33, 18, 27]. McLaughlin *et al*. [21] and Wu *et al*. [30] proposed a basic pipeline for deep video re-ID. First, the frame features are extracted by convolutional neural network. Then a recurrent layer is applied to incorporate temporal context information into each frame. Finally, the temporal average pooling is adopted to obtain video representation. Wu *et al*. [32] further proposed a temporal convolutional subnet to extract local motion information. These methods verify that the temporal information of video can help to identify the person. However, because these methods treat each frame of video equally, the frames with partial occlusion will distort the video representation.

To handle partial occlusion, the attention based approaches are gaining popularity. Zhou *et al*. [43] proposed a RNN temporal attention mechanism to select the most discriminative frames from video. Liu *et al*. [18] used a convolutional subnet to predict quality score for each frame of video. Xu *et al*. [33] presented a Spatial and Temporal Attention Pooling Network, where the spatial attention pooling layer selected discriminative regions from each frame and the temporal attention pooling selected informative frames in the sequence. Similarly, Li *et al*. [15] used multiple spatial attention modules to localize distinctive body parts of person, and pooled these extracted local features across time with temporal attention.

Overall, the above methods process partial occlusion problem by discarding the occluded parts, which results in the loss of spatial and temporal information of video. Different from the existing methods, we explicitly tackle the partial occlusion problem by recovering the occluded parts. Then the recovered parts are leveraged together with the unoccluded parts for robust video reID under partial occlusion.

Image completion. Image completion aims to fill the missing or masked regions in images with plausibly synthesized contents. It has many applications in photo editing, textual synthesis and computational photography. Early works [8, 1] attempted to solve the problem by matching and copying background patches into the missing regions. Recently, deep learning approaches based on Generative Adversarial Network (GAN) [7] had emerged as a promising paradigm for image completion. Pathak *et al*. [23] proposed Context Encoder that generated the contents of an arbitrary image region conditioned on its surroundings. It was trained with pixel-wise reconstruction and an adversarial loss, which produced sharper results than training the model with only reconstruction loss. Iizuka *et al*. [11] improved [23] by using dilated convolution [35] to handle arbitrary resolutions. In [11], global and local discriminators were introduced as adversarial losses. The global discriminator pursued global consistency of the input image, while the local discriminator encouraged the generated parts to be valid. Our proposed STCnet builds on [11] and extends it to exploit the temporal information of video by the proposed temporal attention module. In addition, STCnet employs a guider sub-network endowed with a re-ID cross-entropy loss to preserve the identities of the generated images.

3. Spatial-Temporal Completion network

In this section, we will first illustrate the overview of the proposed STCnet. Then we will demonstrate the details about each module of STCnet. Finally, the objective function to optimize STCnet will be given.

3.1. Network Overview

The key idea of STCnet is to alleviate the interference of occluders on the extracted features for pedestrian retrieval via explicitly recovering the occluded parts with spatio-temporal information of video. The network architecture of STCnet is shown in Figure 1.

STCnet consists of spatial structure generator, temporal attention generator, two discriminators and an ID guider subnetwork. The spatial structure generator leverages the spatial structure of the pedestrian frame, and makes an initial coarse prediction for the contents of occluded parts conditioned on the visible parts of this frame. The temporal attention generator takes use of the temporal patterns of the video, and refines the contents of the occluded parts with the information from adjacent frames. We introduce a local discriminator for the occluded regions to generate more realistic results, and a global discriminator for the entire frame to pursue the global consistency. In addition, an ID guider subnetwork is adopted to preserve the ID label of the frame after completion.

3.2. Spatial Structure Generator

Because of spatial structure of the frames in pedestrian video, the contents of occluded parts can be predicted with the visible parts of the frames. To the end, we design the spatial structure generator to model the correlation between the occluded and visible parts.

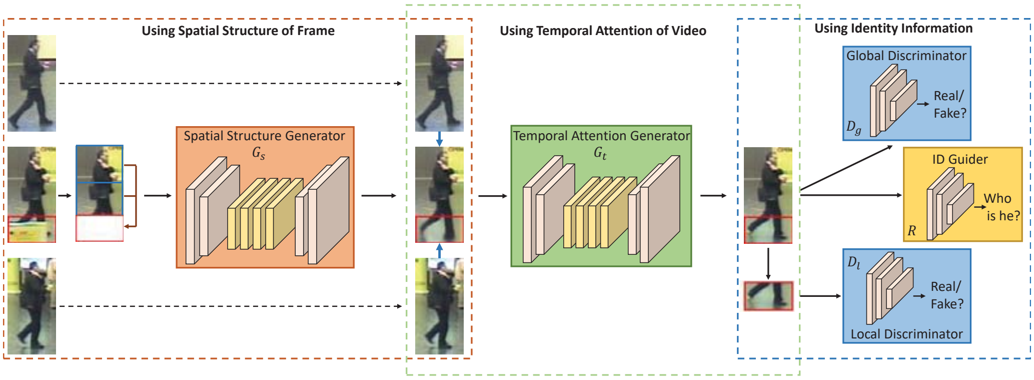

Using Spatial Structure of Frame
Spatial Structure Generator
G_s
Using Temporal Attention of Video
Temporal Attention Generator
G_t
Using Identity Information
Global Discriminator
D_g
Real/ Fake?
ID Guider
R
Who is he?
D_l
Real/ Fake?
Local Discriminator

Figure 1. Overview of STCnet. The spatial structure generator takes the masked frame as input and outputs the generated frame. The temporal attention generator refines the generated frame with the adjacent frames. Two discriminators distinguish the synthesize contents in the mask and whole generated frame as real and fake. The ID guider network is to ensure the identity of the generated frame.

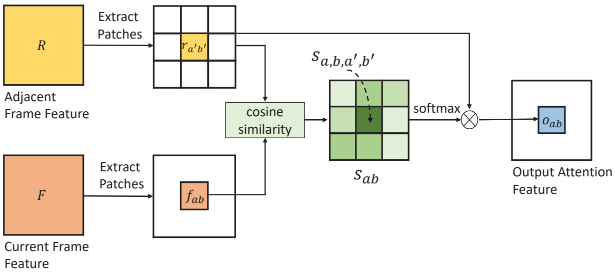


Figure 2. Illustration of the temporal attention layer. For simplicity, we only describe the generation process of one patch ($o_{a,b}$) of output feature. The generation process of other patches is similar.

Spatial structure generator is designed as an autoencoder. The encoder takes a frame with white pixels filled in the occluded parts (all the pixels in the occluded regions are set to 0) as input, which is denoted as masked frame, and produces a latent feature representation of this frame. The decoder takes the feature representation and generates the contents for the occluded parts. In addition, we adopt the dilated convolution [35] in the encoder to enlarge the size of the receptive fields, which can help to propagate the information from distant visible parts to the occluded parts.

The architecture of spatial structure generator is derived from the completion network [11]. In term of layer implementations, we use the convolution with 3 × 3 kernels and ELUs [4] as activation functions. The encoder consists of five convolutional layers and stacks four dilated convolutional layers of that, which decreases the resolution to a quarter of the original size of the input frame. The decoder consists of two deconvolution layers [19] to restore the original resolution of the frame.

3.3. Temporal Attention Generator

In view of the temporal patterns of video, the information from adjacent frames can also be exploited to predict the contents of the occluded parts. So we introduce a novel temporal attention layer, which learns where to attend feature from adjacent frames to generate the contents of the occluded parts. It is differentiable and can be integrated into the temporal attention generator.

The temporal attention layer is able to model relationships between the generated frames of spatial generator and the adjacent frames. For simplicity, we denote the generated frames of spatial generator as current frames. As shown in Figure 2, we first extract patches (3 × 3) in the current frame feature (F) and adjacent frame feature (R). Then, we measure the normalized inner product (cosine similarity) between the patch of F and the patch of R:

$$s_{a,b,a',b'} = \langle \frac{f_{a,b}}{||f_{a,b}||_2}, \frac{r_{a',b'}}{||r_{a',b'}||_2} \rangle, \tag{1}$$

where $f_{a,b}$ denotes the patch centered at location (a, b) in current frame, $r_{a',b'}$ denotes the patch centered at location (a', b') in adjacent frame, $s_{a,b,a',b'}$ indicates similarity between $f_{a,b}$ and $r_{a',b'}$. Then we normalize the similarity with the softmax function:

$$s^*_{a,b,a',b'} = \frac{\exp(s_{a,b,a',b'})}{\sum_{c'd'} \exp(s_{a,b,c',d'})}. \tag{2}$$

Finally, for each patch of current frame, it is updated via aggregating all patches of adjacent frames with weighted summation, where the weighs are decided by the similarity between the corresponding two patches:

$$o_{a,b} = \sum_{a'b'} s^*_{a,b,a',b'} r(a', b'). \tag{3}$$

To integrate temporal attention layer, we introduce three parallel encoders in the temporal attention generator. An encoder for the occluded frame focuses on hallucinating contents, while the other two encoders are for precious and next adjacent unoccluded frame receptively. Two temporal attention layers are appended on top of the encoders to attend on adjacent frames features of interest. Output features from three encoders are then concatenated and fed into a decoder to obtain the final output. The architectures of the encoders and decoder of the temporal generator are the same as those in the spatial generator.

3.4. Discriminator

We adopt a local and a global discriminator to improve the quality of generated contents of the occluded parts. The local discriminator takes the occluded parts as inputs and determines whether the synthesized contents in the occluded parts are real or not. It helps to generate detailed appearance and encourages the generated parts to be valid. The global discriminator takes the entire frames as inputs and regularizes the global structure of the frames. The two discriminators work collaboratively to ensure that the generated contents of occluded parts are not only realistic, but also consistent with surrounding contexts.

The architecture of the two discriminators is similar to [25], which consists of six convolutional layers and a single fully-connected layer. All the convolutional layers use 3×3 kernels and a stride of 2×2 pixels to decrease the frame resolution. The fully-connected layer uses sigmoid as activation function, which outputs the probability that the input is real.

3.5. ID Guider

In order to make the completed (unoccluded) frames boost the person re-ID performance, we introduce an ID guider subnetwork to guide the generators more adapted to re-ID problem. The ID guider subnetwork takes in the completed frames and output the classification results which are forced to be the real categories. In this way, the identity cues of the person are preserved during completion.

We employ ResNet-50 [9] as the backbone network and modify the output dimension of the classification layer to the number of training identities. Following [28], we remove the last spatial down-sampling operation in the ResNet-50 to increase retrieval accuracy with very light computation cost added.

3.6. Object Function

STCnet is trained with three loss functions jointly: a reconstruction loss to capture the overall structure, an adversarial loss to improve the realness, and a guider loss to preserve the ID of the generated frames. Notably, we replace pixels in the non-mask (unoccluded) region of generated frames with original pixels.

We first introduce the reconstruction loss L_r for the spatial generator G_s and temporal generator G_t, which is the L_1 distances between the network output and the original frame:

$$L_r = ||x - \hat{x}_1||_1 + ||x - \hat{x}_2||_1 \tag{4}$$

$$\hat{x}_1 = M \odot G_s((1-M) \odot x) + (1-M) \odot x \tag{5}$$

$$\hat{x}_2 = M \odot G_t(\hat{x}_1, x_p, x_n) + (1-M) \odot x \tag{6}$$

where x is the input of the spatial generator, x_p and x_n are previous and next adjacent frames of x respectively, $\hat{x}_1$ and $\hat{x}_2$ are the predictions of the spatial and temporal generators respectively, M is a binary mask corresponding to the dropped frame region with value 1 wherever a pixel was dropped and 0 for elsewhere, and $\odot$ is the element-wise product operation.

With the global discriminator D_g and local discriminator D_l, we define a global adversarial loss L_{a_1} which reflects the faithfulness of the entire frame, and a local adversarial loss L_{a_2} which reflects the validity of the generated contents in the occluded part:

$$\begin{aligned} L_{a_1} = \min_{G_s,G_t} \max_{D_g} \mathbb{E}_{x\sim p_{data}(x)}[\log D_g(x) \\ + \log D_g(1-\hat{x}_2)] \end{aligned} \tag{7}$$

$$\begin{aligned} L_{a_2} = \min_{G_s,G_t} \max_{D_l} \mathbb{E}_{x\sim p_{data}(x)}[\log D_l(M \odot x) \\ + \log D_l(1-M \odot \hat{x}_2)] \end{aligned} \tag{8}$$

where $P_{data}(x)$ represents the distribution of real frame x.

As for the ID guider network R, the guider loss L_c is the simple cross-entropy loss, which is expressed as:

$$L_c = -\sum_{k=1}^{K} q_k \log R(\hat{x}_2)_k \tag{9}$$

where K is the number of classes and q is the ground truth distribution of the input frames.

Finally, the overall loss function is defined by:

$$L = L_r + \lambda_1(L_{a_1} + L_{a_2}) + \lambda_2 L_c \tag{10}$$

where λ_1 and λ_2 are the weights to balance the effects of different losses.

4. Occlusion-Free Video Person Re-ID

By combining STCnet with a re-ID network, we propose a video re-ID framework VRSTC, which is robust to partial occlusion. The framework of VRSTC is shown in Figure 3. First, a similarity scoring mechanism is proposed to locate the occluded parts of frames. Then, STCnet is adopted to recover the appearance of the occluded parts. Finally, the recovered regions are leveraged with those unoccluded regions to train the re-ID network. Without designing complicated model and loss function, our framework can achieve great performance improvement.

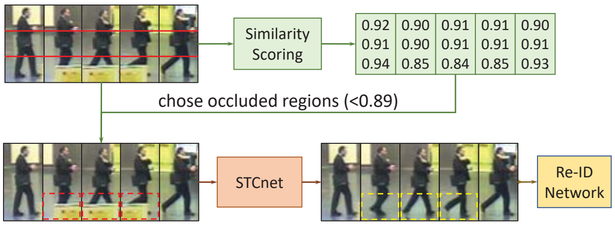


Figure 3. Pipeline of VRSTC.

4.1. Similarity Scoring

The works [18, 43, 33, 3] use the attention mechanism to locate the occluded frames. These approaches usually construct a subnetwork to predict the weight of each frame in video. However, it is difficult for the subnetwork to automatically assign low weights to the occluded frames, as there is no direct supervision for the weights.

Considering the concern above, we propose a similarity scoring mechanism to generate the attention score for each region of frames. Motivated by the observation that the occlusion usually occurs in a few consecutive frames and the occluders have different semantic features from the original body parts, we use the cosine similarity between the frame region feature and the video region feature as the score. Formally, we denote the input video as $I = \{I_t\}_{t=1}^{T}$, where T indicates the length of the video. The frames are vertically divided into three fixed regions equally $I_t = \{I_t^u, I_t^m, I_t^l\}$, where u, m, and l represent the upper, middle and lower part of the frames respectively. The feature representation of each region $\{v_t^k | k \in \{u, m, l\}\}$ is extracted using convolutional neural network. The video region feature is then obtained by average pooling according to temporal domain:

$$\overline{v}^k = \frac{1}{T}\sum_{t=1}^{T} v_t^k, \quad \text{where } k \in \{u, m, l\} \tag{11}$$

Next, the score of each frame region is calculated with the following equation:

$$u_t^k = \left\langle \frac{v_t^k}{||v_t^k||_2}, \frac{\overline{v}^k}{||\overline{v}^k||_2} \right\rangle \tag{12}$$

In the last, we regard those regions with scores lower than a threshold τ (0.89 in our work) as the occluded regions. We replace the occluded regions with the generated regions by STCnet to form a new dataset and train a re-ID network with the new dataset.

4.2. Re-ID Network

Most re-ID networks and loss functions can combine with STCnet. Note that STCnet can combine with the most advanced re-ID models to further enhance the overall performance. In order to verify the effectiveness of STCnet as a kind of data enhancement method, we use a simple re-ID network with average temporal pooling and the cross-entropy loss.

We employ the modified ResNet-50 as the backbone network. In order to capture temporal dependency, we embed the non-local blocks [29] into the re-ID network. Different from the previous works that only build temporal dependency in the end, the non-local blocks can be inserted into the earlier part of deep neural networks. This allows us to build a richer hierarchical temporal dependency that combines both non-local and local information.

5. Experiments

5.1. Datasets and evaluation protocols

iLIDS-VID dataset consists of 600 video sequences, where 300 different identities are captured by two cameras. Each video sequence contains 23 to 192 frames.

MARS dataset is the largest video re-ID benchmark with $1,261$ identities and around $20,000$ video sequences captured from 6 cameras. The bounding boxes are produced by DPM detector [6] and GMMCP tracker [5].

DukeMTMC-VideoReID dataset is a subset of the tracking dataset DuKeMTMC [26] for video person re-ID. The pedestrian images are cropped from the videos for 12 frames every second to generate a tracklet.

Evaluation protocol: We adopt mean Average Precision (mAP) [40] and Cumulative Matching Characteristics (CMC) [2] as evaluation metrics.

5.2. Implementation Details

In this subsection, we give the implementation details of our approach. We use PyTorch [22] for all experiments.

Pre-training a re-ID network. We train ResNet-50 with cross-entropy loss to be the ID guider of STCNet. In training term, four-frame input tracks are cropped out from an input sequence. The frame features are extracted by ResNet-50, then the average temporal pooling is used to obtain the sequence feature. Input images are resized to 256 × 128. The batch size is set to 32. For the data augmentation, we only use random horizontal mirroring for training. We adopt the Adaptive Moment Estimation (Adam) [12] with weight decay of 0.0005. The network is trained for 150 epochs in total, with an initial learning rate of 0.0003 and reduced it with decay rate 0.1 every 50 epochs.

Locating occluded regions. With the pretrained re-ID network as feature extractor, we use the similarity scoring mechanism to generate the score for each frame region. We regard the regions whose scores are lower than τ as the occluded regions, and we define the frames without occluded regions as the unoccluded frames. In our experiment, τ is set to 0.89.

Training STCnet. To train STCnet, we need to build a training set consisting of the input occluded frames and target de-occluded frames. However, there is no ground-truths for the occluded frames. So we only use the unoccluded frames from the training set of target re-ID dataset to train STCnet. Specially, we randomly mask a region of the unoccluded frames as the inputs. The input and target frames are resized to 128×64 and linearly scaled to $[-1, 1]$. The parameters of ID guider are fixed when training STCnet. We optimize the spatial and temporal generators and two discriminators with alternating Adam optimizer, and the learning rate is set to 0.0001. λ_1 and λ_2 are set to 0.001 and 0.1 respectively. Once the training is over, STCnet can recover the appearance of the occluded regions.

Improving re-ID network with de-occluded frames. The occluded regions of the frames in raw re-ID dataset are replaced with the regions generated by STCnet to form a new dataset. Then the re-ID network is trained and tested with the new dataset. We embed the non-local block [29] in the re-ID network to capture temporal dependency of input sequence. According to the experiments in [29], five non-local blocks are inserted to before the last residual block of a stage. Three blocks are inserted into res_4 and two blocks are inserted into res_3, to every other residual block. Other settings are the same as those in the experiments of pre-training a re-ID network. During testing, given an input of entire video, the video feature is extracted using the trained re-ID network for retrieval under cosine distance.

5.3. Ablation Study

5.3.1 Component Analysis of STCnet

We investigate the effect of each component of STCnet by conducting several analytic experiments. Table 1 reports the results of each component of STCnet. Baseline corresponds to ResNet-50 trained on raw target dataset. NL embeds the non-local blocks into the baseline model and improves the results, which indicates that non-local blocks are effective for integrating temporal information of video. In the other experiments of this part, we replace the occluded regions with generated regions by different completion models to form a new dataset and train and test NL on the new dataset.

Spatial structure generator. Spa denotes the spatial structure generator trained only with the spatial reconstruction loss. Compared with NL, Spa improves the rank-1 accuracy by 1.3%, 0.9% and 1.1% on iLIDS-VID, MARS and DukeMTMC-VideoReID respectively. The result shows that the spatial structure generator, which utilizes the spatial information of frames to recover the appearance of occluded regions, is useful for boosting re-ID performance.

Table 1. Comparative analysis of STCnet. The rank-1 CMC accuracy is reported and mAP is reported for MARS and DukeMTMC-VideoReID in brackets.

Methods	iLIDS	MARS	DukeMTMC
baseline	79.8	84.4 (77.2)	91.4 (90.0)
NL	80.1	86.1 (79.9)	91.8 (91.2)
Spa	81.4	87.0 (81.0)	92.9 (92.0)
Spa+AE	81.3	87.0 (80.8)	92.9 (91.9)
Spa+TAE	81.9	87.3 (81.0)	93.2 (92.2)
Spa+Tem	82.5	87.8 (81.6)	93.8 (92.7)
Spa+Tem+LD	82.7	87.9 (81.7)	94.1 (92.8)
Spa+Tem+LD+GD	82.9	87.9 (81.9)	94.4 (93.0)
STCnet	**83.4**	**88.5 (82.3)**	**95.0 (93.5)**

Temporal attention generator. Spa+Tem consists of spatial and temporal generators, which is trained with both spatial and temporal reconstruction loss. By comparing Spa and Spa+Tem, we can see that the proposed temporal generator further improves accuracy. We argue that the temporal attention layer can attend the information from adjacent frames, which makes the generated frames more semantically consistent with the video sequence. The re-ID network (NL) can then extract better temporal information of the resulting sequence, leading to a more discriminative video feature representation.

It is noteworthy that the improvement of temporal generator does not come from the increased depth by naively adding extra layers to the spatial generator. To see this, we also try two variants of temporal generator: **Autoencoder (AE)** and **Temporal Autoencoder (TAE)**. AE is a standard autoencoder and only takes the predictions of spatial generator as inputs. It has the same encoder and decoder with temporal generator expect the number of filters in the encoder is tripled. This controls for the total number of parameters in AE compared to temporal generator. TAE is the temporal generator without the temporal attention layer. As shown in Table 1, Spa+AE does not increase the accuracy compared to Spa. This shows that the improvement of temporal generator is not because it adds extra layers to the spatial generator. In addition, the temporal generator performs better than TAE. This improvement shows that the proposed temporal attention layer makes better use of the temporal information to generate more discriminative frames.

Discriminators. Spa+Tem+LD consists of the two generators and the local discriminator. Spa+Tem+LD+GD further incorporates the global discriminator. Both are trained with reconstruction and adversarial losses. From the results, we can see that the discriminators only slightly improve the performance. We argue that the discriminators aim to generate more visually realistic frames, without bringing too much additional discriminant information for re-ID.

ID guider network. The final model STCnet is trained with the reconstruction, adversarial and guider losses. The generated samples achieve better performance with the ID guider, which suggests that the ID guider is beneficial to generate suitable samples for training re-ID network. The improvement can be attributed to the ability of preserving

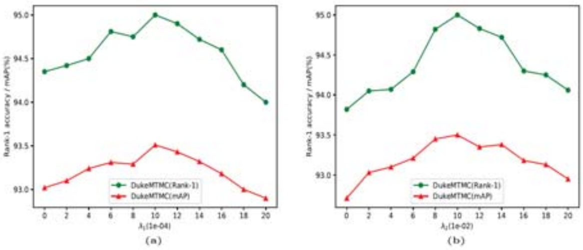


Figure 4. The rank-1 and mAP on DukeMTMC-VideoReID (a) different λ_1 and fixed λ_2=0.1, (b) different λ_2 and fixed λ_1=0.001.

Table 2. Comparison of different threshold τ in the similarity scoring mechanism. The rank-1 CMC accuracies are reported and mAP are reported for MARS and DukeMTMC-VideoReID in brackets.

threshold (τ)	iLIDS	MARS	DukeMTMC
0 (baseline)	79.8	84.4 (77.2)	91.4 (90.0)
0.88	80.0	84.8 (77.2)	91.5 (90.3)
0.89	**80.3**	**84.9 (77.4)**	**91.7** (90.5)
0.91	78.8	84.2 (77.0)	91.4 (**90.6**)
0.93	78.3	83.6 (76.6)	91.4 (90.5)

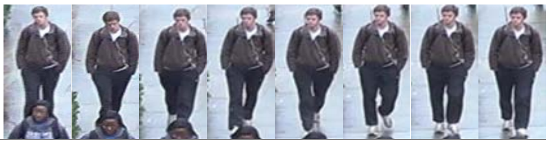
Images								
up-part	0.961	0.954	0.966	0.963	0.967	0.958	0.956	0.959
mid-part	0.971	0.970	0.974	0.968	0.972	0.971	0.965	0.969
low-part	0.862	0.873	0.881	0.887	0.901	0.911	0.924	0.928

Figure 5. Scores of similarity scoring mechanism from one sequence. Red represents small score.

the underlying visual cues associated with the ID labels.

5.3.2 Influence of the parameters λ_1 and λ_2

λ_1 and λ_2 are two parameters to balance the relative effects of the adversarial loss and guider loss respectively. We analyze the impact of the λ_1 and λ_2 on DukeMTMC-VideoReID, and the results are shown in Figure 4 (a) and (b) respectively. We observe that our method achieves the best performance when λ_1 is set to 0.001 and λ_2 is set to 0.1. Notice that there will be a big performance degradation when λ_1 or λ_2 are too big. The main reason is that STC-net becomes difficult to converge if the adversarial loss or guider loss takes a dominant role.

5.3.3 Influence of the threshold τ

We also carry out experiments to investigate the effect of varying the threshold τ in the similarity scoring mechanism. The experiment setting is as follows. Giving an input video sequence, we first discard the frames with occluded regions whose scores are lower than τ. The video feature is then obtained with the remaining frames using average temporal pooling. Finally, we use the obtained video feature to compute the similarity between videos under cosine distance. Notably, when $\tau = 0$, we keep all frames of a video, which is the same as the baseline model.

Table 3. Comparison with related methods on MARS. * denotes those requiring optical flow as inputs.

Methods	rank-1	rank-5	rank-10	mAP
Mars [39]	68.3	82.6	89.4	49.3
SeeForest [43]	70.6	90.0	97.6	50.7
Seq-Decision [36]	71.2	85.7	91.8	-
Latent Parts [14]	71.8	86.6	93.0	56.1
QAN [18]	73.7	84.9	91.6	51.7
K-reciprocal [42]	73.9	-	-	68.5
RQEN [27]	77.8	88.8	94.3	71.7
TriNet [10]	79.8	91.4	-	67.7
EUG [31]	80.8	92.1	96.1	67.4
STAN [15]	82.3	-	-	65.8
Snipped [3]	81.2	92.1	-	69.4
Snippet+OF* [3]	86.3	94.7	**98.2**	76.1
VRSTC	**88.5**	**96.5**	97.4	**82.3**

As shown in Table 2, there is an improvement in performance when τ is increased, which implies that the discarded frames would have corrupted the representation of video. This result implicitly demonstrates the scores achieved by similarity scoring can locate occluded frames. However, as τ is further increased, the accuracy drops gradually. The main reason is that the unoccluded frames may be discarded with too large threshold. The network achieves the best performance when $\tau = 0.89$. So we set τ to 0.89 in our experiments. Notably, the introduction of the frames completed by STCnet can further improve the performance (see Table 1), which demonstrates that the contents restored by STCnet can help identify the person.

In order to demonstrate the similarity scoring mechanism more intuitively, the scores of one sequence from DukeMTMC-VideoReID is visualized in Figure 5. Due to the occlusion from another person, the scores of the lower parts of the first four frames are relatively small. This results further demonstrate the score achieved by the proposed similarity scoring mechanism can reflect the visibility of each region.

5.4. 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 MARS, DukeMTMC-VideoReID and iLIDS-VID, respectively. On MARS and DukeMTMC-VideoReID, our approach outperforms the best existing methods. We attribute the improvements to

Table 4. Comparison with methods on DukeMTMC-VideoReID.

Methods	rank-1	rank-5	rank-10	mAP
EUG [31]	83.6	94.6	97.6	78.3
VRSTC	**95.0**	**99.1**	**99.4**	**93.5**

Table 5. Comparison with related methods on iLIDS-VID.

Methods	rank-1	rank-5	rank-10	rank-20
LFDA [24]	32.9	68.5	82.2	92.6
KISSME [13]	36.5	67.8	78.8	87.1
LADF [16]	39.0	76.8	89.0	96.8
STFV3D [17]	44.3	71.7	83.7	91.7
TDL [34]	56.3	87.6	95.6	98.3
Mars [39]	53.0	81.4	-	95.1
SeeForest [43]	55.2	86.5	-	97.0
CNN+RNN* [21]	58.0	84.0	91.0	96.0
Seq-Decision [36]	60.2	84.7	91.7	95.2
ASTPN* [33]	62.0	86.0	94.0	98.0
QAN [18]	68.0	86.8	95.4	97.4
RQEN [27]	77.1	93.2	97.7	99.4
STAN [15]	80.2	-	-	-
Snippet [3]	79.8	91.8	-	-
Snippet+OF* [3]	**85.4**	**96.7**	**98.8**	**99.5**
VRSTC	83.4	95.5	97.7	**99.5**

the recovered contents of the occluded parts. The effective combination with STCnet makes our approach superior than the methods which only use the raw dataset. It is worth noting that DukeMTMC-VideoReID is recently proposed by [31] and our baseline model has outperformed [31] by 7.8% and 11.7% on rank-1 and mAP respectively. We hope it will serve as a new baseline on DukeMTMC-VideoReID. On iLIDS-VID, our approach achieves slightly lower performance than Snippet+OF [3]. Note that Snipper+OF uses additional optical flow as input to provide motion features, which is not utilized in our framework. In addition, our approach outperforms Snippet (without optical flow) significantly, which is a more fair comparison.

5.5. Visualizing the effect of STCnet

We visualize the generated frames of STCnet for intuitive exploration. Some partially occluded images are selected for evaluation. Figure 6 provides a vivid illustration how STCnet recovers the contents of occluded parts and improves the extracted features. Specifically, when a person is occluded by some body part of other pedestrians, the feature representation extracted for the person is often corrupted by the visual appearances of the other pedestrians. As shown in the sixth column of Figure 6 (c), the part of other pedestrians is activated by the re-ID network, which harms the feature representation of the target person. In addition, when a person is occluded by the environmental objects such as indicator and bicycles, there will be severe loss of body information in the feature extracted from the person (*e.g.* the second column of Figure 6 (c)). On the contrary, once STCnet recovers the contents of the occluded regions, the re-ID model will take more effective regions into account and discover new discriminative clues therefrom to recognize the person more correctly.

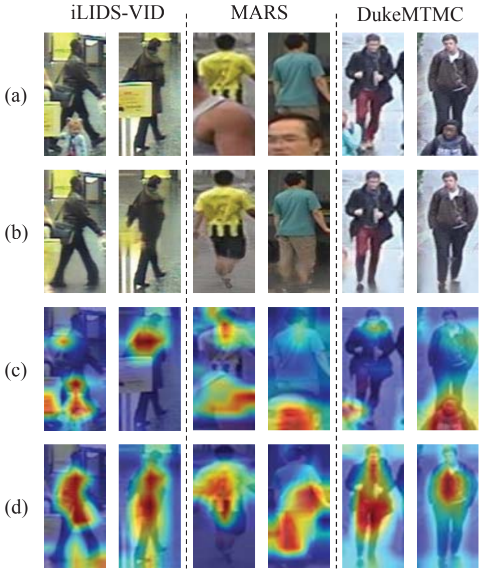


Figure 6. Visual examples of STCnet. From top to bottom: (a) original image, (b) output of STCnet, (c) the activation maps of original image (d) the activation maps of completed image. Warmer color with higher value

6. Conclusion

In this work, we present a novel framework combined a re-ID network with a completion network STCnet for video re-ID under partial occlusion. Aiming at explicitly tackling the partial occlusion problem, we design the STCnet to recover the appearance for the occluded regions and leverage the recovered regions with the unoccluded regions to train the re-ID network. Experiments on three datasets show that the proposed method outperforms the state-of-the-art video re-ID approaches.

In the future, we will explore other types of deep generative architectures for recovering the appearance for the frames with extremely severe occlusion.

Acknowledgement This work is partially supported by National Key R&D Program of China (No.2017YFA0700800), Natural Science Foundation of China (NSFC): 61876171 and 61572465, and Beijing Municipal Science and Technology Program: Z181100003918012.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):24, 2009.

[2] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the roc curve and the cmc. In *AUTOID*, pages 15–20, 2005.

[3] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018.

[4] D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units. *arXiv preprint arXiv:1511.07289*, 2015.

[5] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, pages 4091–4099, 2015.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[7] I. Goodfellowa, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[8] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26(3):4, 2007.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

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

[11] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.

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

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

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

[15] S. Li, S. Bak, P. Carr, C. Hetang, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[16] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610–3617, 2013.

[17] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatiotemporal appearance representation for video-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015.

[18] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 4694–4703, 2017.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[20] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(4):3656–3670, 2014.

[21] N. McLaughlin, J. M. del Rincon, and P. C. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS workshop*, 2017.

[23] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[24] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian reidentification. In *CVPR*, pages 3318–3325, 2013.

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multitarget, multicamera tracking. In *ECCV Workshop*, 2016.

[27] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. *arXiv preprint arXiv:1711.08766*, 2017.

[28] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[29] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[30] L. Wu, C. Shen, and A. V. D. Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016.

[31] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Quyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018.

[32] Yang Wu, Jie Qiu, Jun Takamatsu, and Tsukasa Ogasawara. Temporal-enhanced convolutional network for person re-identification. In *AAAI*, pages 7412–7419, 2018.

[33] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4743–4752, 2017.

[34] J. You, A. Wu, X. Li, and W. Zheng. Top-push video-based person re-identification. In *CVPR*, pages 1345–1353, 2016.

[35] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[36] J. Zhang, N. Wang, and L. Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, pages 6781–6789, 2018.

[37] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016.

[38] 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 *CVPR*, pages 1077–1085, 2017.

[39] 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 *ECCV*, pages 868–884, 2016.

[40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[41] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011.

[42] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661, 2017.

[43] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785, 2017.