Knowledge-Based Systems 249 (2022) 108884

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

SFGN: Representing the sequence with one super frame for video person re-identification

Xiao Pan [a], Hao Luo [a], Wei Jiang [a,*], Jianming Zhang [a], Jianyang Gu [a], Peike Li [b]

[a] Zhejiang University, Hangzhou, China
[b] University of Technology Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 1 December 2021
Received in revised form 19 April 2022
Accepted 20 April 2022
Available online 5 May 2022

Keywords:
Person re-identification
Video-based
Super frame
Deep learning

ABSTRACT

Video-based person re-identification (V-Re-ID) is more robust than image-based person re-identification (I-Re-ID) because of the additional temporal information. However, the high storage overhead of video sequences largely stems the applications of V-Re-ID. To reduce the storage overhead, we propose to represent each video sequence with only one frame. However, directly picking one frame from each sequence will reduce the performance dramatically. Thus, we propose a brand-new framework called super frame generation network (SFGN), which can encode the spatial–temporal information of a video sequence into a generated frame, which is called "super frame" to distinguish from the directly picked "key frame". To achieve super frames of high visual quality and representation ability, we carefully design the specific-frame-feature fused skip-connection generator (SFSG). SFSG takes the role of a feature encoder and the co-trained image model can be seen as the corresponding feature decoder. To reduce the information loss in the encoding–decoding process, we further propose the feature recovery loss (FRL). To the best of our knowledge, we are the first to propose and relieve this issue. Extensive experiments on Mars, iLIDS-VID, and PRID2011 show that the proposed SFGN can generate super frames of high visual quality and representation ability. For the code, please visit the project website: https://github.com/pansanity666/SFGN_VideoReID/tree/main.

1. Introduction

Person re-identification (Re-ID) aims to retrieve a specific person across different cameras and is becoming increasingly important because of its extensive applications in surveillance [1] and tracking [2]. Re-ID is quite challenging due to the complicated environment (e.g., cluttered background, occlusions, and different lighting conditions) in realistic scenes. Recently, the community has achieved great progress with the development of deep learning.

Generally, person Re-ID can be divided into image-based person Re-ID (I-Re-ID) [3–9] and video-based person Re-ID (V-Re-ID) [10–12] based on the different input types. I-Re-ID takes the images as input, which means it extracts frame-level features from every single frame. By contrast, V-Re-ID takes the video sequences as input, which means it first extracts frame-level features and then aggregates them to obtain the final sequence-level feature.

Previous works of V-Re-ID mainly focused on the *aggregation* of frame-level features, such as 3DCNN [13], RNN [14–16], and attention [10,11,17–21]. The aggregated sequence-level features are inherently more robust and discriminative than frame-level features because of the spatial–temporal information in video sequences.

However, the high **storage overhead** of video sequences may largely stem the applications of V-Re-ID. For instance, suppose a city-scale surveillance system that may generate 100M sequences on average per day. For each sequence, assume that 100 frames are stored. Given that the size of each frame (a JPG image with a resolution of 128×64) is around 3 (kByte), the overall storage overhead increased per day will be around 27.9 (TB) $\approx 100 * 100M/(3 * 1024^3)$. Notably, the sequence number, sequence length, and image size may be far beyond this assumption in practice.

To reduce the storage overhead, an ideal solution is to use only one frame to represent the whole sequence without losing performance. Nevertheless, directly picking one key frame from the sequence will decrease the performance dramatically, as illustrated in Table 3. This motivates us to generate such key frame from the sequence-level feature since it contains the information of the entire video sequence.

* Corresponding author.
E-mail addresses: xiaopan@zju.edu.cn (X. Pan), haoluocsc@zju.edu.cn (H. Luo), jiangwei_zju@zju.edu.cn (W. Jiang), ncsl@zju.edu.cn (J. Zhang), gujianyang@zju.edu.cn (J. Gu), peike.li@student.uts.edu.au (P. Li).

https://doi.org/10.1016/j.knosys.2022.108884
0950-7051/

However, if we send the sequence-level features into a generator directly, the best we can get are average-looking images (see Mid+Skip and Avg+Skip in Fig. 3). These low-quality images are not clear enough for the users (e.g., police) to verify the retrieval results. This leads us to change the paradigm from "directly generating" to "encoding". Specifically, instead of generating from the sequence-level features directly, we propose to encode them into clean images via a designed generator (encoder).

To achieve this goal, we propose a novel framework called super frame generation network (**SFGN**), which can encode a sequence-level feature into one high-quality **"super frame"**.[1] SFGN is composed of a video model, an image model, and an elaborately designed *specific-frame-feature fused skip-connection generator (SFSG)*, as illustrated in Fig. 2. The video model plays the role of a feature provider, which provides the discriminative sequence-level features for encoding and supervision. SFSG plays the role of a feature encoder, which encodes the sequence-level feature of a sequence into a high-quality super frame. The co-trained image model can be seen as the corresponding feature decoder, which decodes the sequence-level feature from a super frame. To reduce the information loss in the encoding–decoding process, we further propose the *feature recovery loss (FRL)*.

The comparison of pipelines between SFGN and other methods is illustrated in Fig. 1. The incoming video sequences are embedded into super frames, which are then stored in the gallery instead of the original sequences. When online applications are needed, the sequence-level features can be decoded from super frames by the co-trained image model. With SFGN, we achieve a **trade-off** between the computation overhead and the storage overhead. Nevertheless, we believe that the increased computation overhead is acceptable compared with the expensive storage overhead in edge devices since the training and encoding processes can be performed offline, and the decoding process is fast with GPU.

The experimental results on Mars, iLIDS-VID, and PRID2011 show that the generated super frames are of high visual quality, and can achieve a performance similar to those of sequence-level features.

The contributions of this work are summarized as follows:

- We put forward the storage overhead issue for V-Re-ID and relieve it with the proposed novel framework called SFGN, which is composed of a video model (feature provider), a generator (feature encoder), and an image model (feature decoder).
- We design the SFSG elaborately to achieve high-quality super frames and propose the FRL to reduce information loss.
- We conduct extensive experiments on Mars, iLIDS-VID, and PRID2011 to prove the effectiveness of SFGN.

2. Related works

2.1. Video-based person Re-ID

Sampling strategy. The sampling strategy is an important issue for V-Re-ID considering that there are multiple frames in a video sequence.

In the training stage, random strategy [10,17,18,22], is mostly used because it can provide large number of different combinations. In addition, several works attempt to select key frames from the original sequences during training. Wang et al. [23] and Zhang et al. [24] proposed leveraging the gait information to sample several key frames. They leveraged the flow energy profile

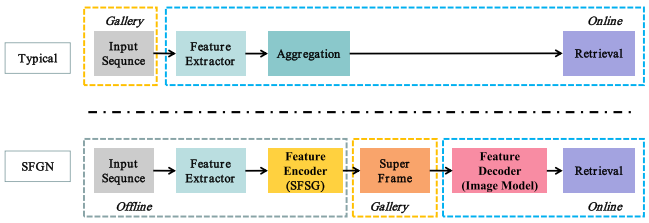


Fig. 1. The comparison of the pipelines in practical applications (**inference stage**) between our proposed SFGN and the typical methods. The typical methods store sequences in the gallery, while SFGN only needs to store one super frame per sequence.

(FEP) [23] to determine the period of walking cycle. However, FEP requires a clean background and is sensitive to occlusion. For a dataset, such as Mars, which has relatively low-quality cropped pedestrian images, obtaining a walking cycle will be difficult [24]. Song et al. [15] proposed using LOMO [25] to select key frame groups based on the distance to the average LOMO feature center of randomly chosen frames. In this work, we utilize the random strategy in the training stage because it is both concise and powerful.

In the inference stage, dense strategy, which averages all the frame-level features in a sequence as in [14,22,26], is mostly used, because it can make full use of the information in the video sequences. Dense strategy provides outstanding results in the inference stage, but it also requires a large amount of storage space. Thus, we elaborately designed the SFGN to reduce the storage overhead by generating powerful super frames as the representation of video sequences. The generated super frames are encoded with the discriminative sequence-level features. Using SFGN, only one super frame is stored per sequence while the high performance is reserved.

Aggregation methods. The works on V-Re-ID mainly focus on improved aggregation methods to leverage spatial–temporal cues better. The mostly researched aggregation methods are mainly 3DCNN-based [13], RNN (LSTM)-based [14–16], and attention-based [10,11,17,19–21].

Li et al. [13] proposed a two-stream network, one for spacial and another for temporal learning. The temporal stream is constructed by inserting several multi-scale 3D convolution layers into 2D CNN. Low computational efficiency is the largest problem of 3DCNN.

Several works used RNN to aggregate the frame-level features due to its satisfactory performance in sequence modeling. Song et al. [15] adopted the Bi-LSTM to extract the global and local features and then combine them for discrimination. Liu et al. [14] introduced a refining recurrent unit (RRU) to recover the missing part in the sequence and then used a spatial–temporal clues integration module (STIM) to obtain the spatial–temporal features. However, Zhang et al. [16] found that RNN was not so good at modeling temporal information as we expected, and they proposed an orderless method to address the V-Re-ID problem.

Attention mechanism is also extensively researched in V-Re-ID. The main purpose of attention is to provide different weights to different frames or fine-grained areas. Liu et al. [27] proposed a quality-aware network (QAN) to estimate quality scores for frames in a sequence and fused them weighted by the estimated scores. Wang et al. [19] proposed feature learning attention (FLA) and feature fusion attention (FFA) based on Gaussian function. Jiang et al. [11] considered the temporal attention together with semantic attention and generated temporal attention from intra-frame and inter-frame aspects. Zhang et al. [10] proposed multi-granularity referenced attentive feature aggregation (MG-RAFA) to generate multi-granularity attention with the help of a global

[1] We term it as "super frame" because it is powerful enough to reserve the information of the whole sequence.

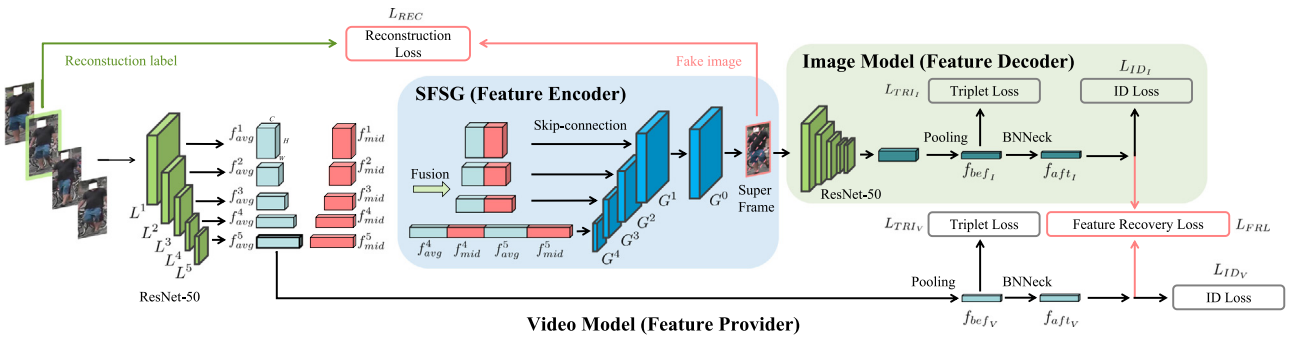


Fig. 2. The framework of our SFGN. It is composed of a video model (V-BoT), an image model and a SFSG model. The video model provides the discriminative sequence-level features for encoding and supervision, SFSG encodes the sequence-level features into super frames, and the co-trained image model decodes the sequence-level features from super frames. In the training stage, we randomly pick T frames from each sequence as in [10,17,18,22]. We set T = 4 as an example here. In the inference stage, all the frames in a sequence are used (dense strategy). During applications, the incoming video sequences are embedded into super frames, which are then stored in the gallery. When online retrieval is needed, the sequence-level features can be decoded from super frames by the image model.

reference. Eom et al. [20] leveraged spatial and temporal memories to refine frame-level features and then aggregated them into sequence-level features, of which the former stores frequently emerged distractors and the latter saves attentions. Hou et al. [21] took two branches of frames with different resolutions as input and each branch contained multiple parallel and diverse attention modules.

In this work, we surprisingly find that using the simple temporal average pooling as the aggregation model can already achieve performance on par with these SOTA methods.

2.2. Image generation in Re-ID

Many works attempted to leverage the generative adversarial network (GAN) [28] to solve the problem of data augmentation and domain adaption for Re-ID with the success of GAN in recent years.

Data augmentation. Zheng et al. [29] first introduced DCGAN [30] to generate the unlabeled image from random noises for data augmentation. Hou et al. [31] proposed STCnet to recover the appearance of the occluded parts in the video sequence. Qian et al. [32] proposed PN-GAN to generate person images conditioned on the normalized pose. Similar works, such as [33–35], also leveraged the pose information for generation. However, these methods isolated the generative model and discriminative models. Zheng et al. [36] proposed DG-Net to couple the Re-ID learning and image generation in an end-to-end manner.

Domain adaption. Several works [37,38] leveraged the image generation model to solve the problem of domain adaption for Re-ID. Based on Cycle-GAN [39], Wei et al. [37] proposed PTGAN to transfer the images from the source domain to the target domain, which bridged the domain gap between datasets. Wei et al. [38] integrated a SiaNet with CycleGAN to constrain the self-similarity and domain-dissimilarity better.

Different from the works mentioned above, our proposed SFGN targets on generating a high-quality and informative super frame to reduce the storage overhead for V-Re-ID. To the best of our knowledge, we are the first to propose this issue.

3. Proposed method

We propose a novel framework called SFGN, which can encode the information of the whole sequence into a high-quality super frame, to reduce the storage overhead for V-Re-ID. As shown in Fig. 2, our proposed SFGN is composed of a video model, a carefully designed SFSG, and an image model. We will introduce the details of each part in the following paragraphs.

3.1. Video model

Similar to those in previous works [10,17,18,22], the video model is composed of a feature extractor and a feature aggregation model. In this work, we use ResNet-50 as the feature extractor and the simple temporal average pooling as the aggregation model by default. Similar to BoT [5,6], which is a strong baseline for I-Re-ID, we add a BNNeck after the aggregated sequence-level features. We borrow most of the training tricks from [5,6], except that we change the frame-level REA to sequence-level synchronize REA [10,40,41]. We name this simple baseline model for V-Re-ID as *V-BoT*.

Formally, the video model takes T frames from a sequence as input, which can be represented as $\{F_i\}_{i=1}^{T}$. There are totally five layers in ResNet-50, namely, one convolution layer and four bottlenecks, which can be represented as $\{L^j\}_{j=1}^{5}$. We represent the frame-level features of F_i after different layers as $\{f_i^j\}_{j=1}^{5}$, where j represents for the jth layer, and $f_i^j \in \mathbb{R}^{H_j \times W_j \times C_j}$ (H_j, W_j and C_j represent for the height, width, and the channel of the feature map after the jth layer). Then, we represent the sequence-level feature after jth layer as $f_{avg}^j \in \mathbb{R}^{H_j \times W_j \times C_j}$, which is calculated by the temporal average pooling:

$$f_{avg}^j = \frac{1}{T}\sum_{i=1}^{T} f_i^j. \tag{1}$$

Next, we conduct spatial average pooling on f_{avg}^5 to get $f_{bef_V} \in \mathbb{R}^{2048}$. Finally, f_{bef_V} is sent to a batch normalization neck (BNNeck) [5,6] to get $f_{aft_V} \in \mathbb{R}^{2048}$. To make the sequence-level features discriminative, as in [5,6], we send f_{bef_V} to the hard mining triplet loss [42] and f_{aft_V} to ID loss [4].

Surprisingly, we find the simple baseline model V-BoT can already achieve quite comparable performance with other SOTA methods who designed sophisticated aggregation models [10,11,17–21] (see Section 4.3), and can serve as a strong baseline for V-Re-ID community.

3.2. Specific-frame-feature fused skip-connection generator

We propose the SFSG to obtain high-quality super frames encoded with sequence-level features. Constrained by the losses, f_{bef_V} and f_{aft_V} are the most discriminative sequence-level features. However, they are lack of the important spatial information due to the spatial pooling operation. Therefore, we use f_{avg}^5 (the feature before spatial pooling) for up-sampling considering that it is both discriminative and contains spatial information.

Nevertheless, directly using f_{avg}^5 for reconstruction can only generate **blurry** images (see Mid and Avg in Fig. 3). The blur

is caused by the loss of fine-grained information during the down-sampling stage in ResNet-50. As we all know, the shallow-level features of ResNet-50 contain rich fine-grained information. Therefore, inspired by U-Net [43], we propose to **skip-connect** the shallow-level features during the up-sampling stage. The skip-connection can compensate for the missing fine-grained information.

After adding the skip-connection, the images become clearer, but still **average-looking** (see Mid+Skip and Avg+Skip in Fig. 3). In fact, this is intrinsically caused by the aggregation operation when getting sequence-level features. Sequence-level features are the aggregation of frame-level features, thus, they are intrinsically similar to the pixel-level averaged images after several steps of up-sampling, even when supervised by a clean image (see Mid+Skip in Fig. 3). Intuitively, if we want to get a clean image, the feature used for the reconstruction should contain a clean feature (i.e., a feature of a single frame instead of an aggregated feature). At the same time, we still need the feature for reconstruction to contain the sequence-level feature because we want the generated super frame to maintain the information of the entire sequence. This motivates us to **fuse** the sequence-level feature with a clean frame-level feature in the up-sampling stage. By doing so, the paradigm has changed from "directly generating" to "encoding". The generated super frame is visually similar to the assigned frame who provides the clean frame-level feature, but encoded with the discriminative sequence-level feature. The experimental results show that the performance is irrelevant to the quality of the assigned frame (see Section 4.5.1). Thus, we call it "the specific frame". For simplicity, we use the frame in the **middle** of the sequence by default.

Formally, we represent the frame in the middle of the sequence as F_{mid}. Considering that we set the last stride of ResNet-50 as 1 [5,6], f_i^4 and f_i^5 share the same feature size ($H_4 = H_5 = 16, W_4 = W_5 = 8$) but have different channel numbers ($C_4 = 1024, C_5 = 2048$). To make full use of the existing features, we concatenate them as the first layer input of SFSG. We represent it as $[f_{avg}^4, f_{mid}^4, f_{avg}^5, f_{mid}^5]$, where $[\cdot,\cdot]$ represents the concatenation operation. SFSG contains five layers, which can be represented as $\{G^j\}_{j=0}^4$, where G^j represents the jth layer. For clarity and symmetry (see Fig. 2), we begin the superscript from 0. G^4 represents the first (input) layer, and G^0 represents the last (output) layer. We obtain the input of each layer I^j as follows:

$$I^j = \begin{cases} G^1(I^1), & j = 0 \\ \left[G^{j+1}(I^{j+1}), f_{avg}^j, f_{mid}^j\right], & j \in \{1,2,3\} \\ \left[f_{avg}^4, f_{mid}^4, f_{avg}^5, f_{mid}^5\right], & j = 4 \end{cases}. \tag{2}$$

The detailed structure of each layer will be illustrated in Section 4.2.

We utilize the L_1 reconstruction loss to supervise the reconstruction, which is a common practice for the reconstruction of images [31,32,36,44]:

$$L_{REC} = \|\hat{F}_{sf} - F_{mid}\|_1, \tag{3}$$

where $\hat{F}_{sf}$ represents the output of SFSG, and F_{mid} represents the original middle frame.

3.3. Image model

The SFSG can be seen as a feature encoder, and the image model takes the role of a corresponding feature decoder. Using SFSG, the sequence-level feature of a video sequence is encoded into a super frame. Thus, the remaining problem is how to decode the sequence-level feature in a super frame.

Intuitively, the generated super frames share the same semantic structure as the original frames. Thus, we use the same model structure as the video model except the aggregation part since we only have one frame per sequence here. Similar to the video model, we use hard mining triplet loss [42] and ID loss [4] to ensure that the decoded features are discriminative.

3.4. Feature recovery loss

Although the hard mining triplet loss [42] and ID loss [4] in the image model can make the decoded features discriminative to some extent, they are not **explicit** enough to ensure that the decoded features are as discriminative as the encoded sequence-level features. Thus, we further propose FRL to promote the decoding of sequence-level features by minimizing the distance between the decoded features and the original sequence-level features. Intuitively, our goal is to make the discriminative ability of the decoded features and the encoded sequence-level features as similar as possible, that is, as close as possible. As such, minimizing the distance between them is natural and reasonable.

Formally, the feature after BNNeck in the image model is represented as f_{aft_I}, and that of the video model is represented as f_{aft_V}. FRL L_{FRL} is calculated as follows:

$$L_{FRL} = \|f_{aft_I} - f_{aft_V}\|_1. \tag{4}$$

Note that we opt to constrain the ones after the BNNeck instead of the ones before because the former achieve higher performance, as mentioned in [5,6]. Our experimental results in Section 4.5.2 show that L_1 and L_2 distance exhibit similar performance. For simplicity, we use L_1 distance by default.

3.5. Overall losses

The overall loss function is defined by:

$$L = L_{TRI_V} + L_{ID_V} + L_{TRI_I} + L_{ID_I} + L_{REC} + L_{FRL}, \tag{5}$$

where L_{TRI_V} and L_{ID_V} represent for the hard mining triplet loss [42] and ID loss [4] for video model, separately; and L_{TRI_I} and L_{ID_I} represent for the ones for the image model, respectively.

4. Experimental results

4.1. Datasets and evaluation metrics

We conduct our experiments on Mars [45], iLIDS-VID [23], and PRID2011 [46].

Mars contains 1261 persons with around 20,000 video sequences from 6 cameras, which are captured on the campus of Tsinghua University. The bounding boxes are cropped by DPM detector [47] and GMMCP tracker [48], which leads to the relatively low quality of frames (false detection/tracking). Mars is the most challenging and persuasive dataset so far due to its large scale, complicated environment, and unstable frame quality.

iLIDS-VID is shot in an airport hall. It contains 300 persons with 600 sequences, 2 sequences from 2 different cameras per person. The sequence length ranges from 23 to 192. The bounding boxes are cropped manually.

PRID2011 is captured in an uncrowded outdoor environment with 2 cameras, and the background is relatively clean. It has 400 sequences for 200 persons who appear in both cameras. The sequence length ranges from 5 to 675. Following the setting of [23], sequences with more than 21 frames from 178 persons are used in our experiments.

For Mars, we follow the train/test protocol in [45] and evaluate the performance using mean average precision (mAP) and the cumulative match characteristic (CMC). For iLIDS-VID and PRID2011, as in [23], we randomly split the dataset into two subsets of equal size for 10 times and then average their CMC performance.

4.2. Implementation details

We pre-train the ResNet-50 [49] on ImageNet [50]. For each batch, we utilize the *PK* strategy in [42], namely, P identities per batch, and K sequences per identity. We set P as 8 and K as 4. The sequence length T for training is set as 8 for Mars, and 6 for iLIDS-VID and PRID2011. In the inference stage, we report the performance of dense strategy unless otherwise specified. The features after BNNeck are used for retrieval with cosine distance, except for several splits of iLIDS-VID and PRID2011, where we find that features before BNNeck gives better performance. This is reasonable since these two datasets are relatively small for deep-based methods. Therefore, their performance is not so stable as that on Mars. All the performances illustrated here are the ones without re-ranking.

We implement the model with Pytorch. Adaptive moment estimation (Adam) [51] is used as the optimizer. To accelerate the convergence of SFSG, we detach SFSG when backward propagating the gradient of reconstruction loss. Experiments are conducted on 4 GTX 1080ti. We train SFGN and V-BoT on Mars for approximately 13 and 11 h, respectively.

The detailed structure of SFSG is illustrated in Table 1. We refer readers to our released code for more implementation details.

Table 1
The detailed structure of SFSG. $G^4 - G^1$ are de-convolution layers, and G^0 is a convolution layer. "RefPad" represents for reflection padding. The shapes are represented in the form of (H,W,C). The channels of input shape are represented by "×2" because the inputs are the fusion of sequence-level features and middle frame features. "k", "s", "p", and "c" represent for "kernel size", "stride", "padding", and "channel", separately. "IN" represents for instance normalization.

Layer	Input shape	Output shape	Name	Parameters
G^4	18,6,3072 × 2	32,16,1024	Deconv IN ReLU	k = 3, s = 2, p = 1 c = 1024 –
G^3	32,16,1024 × 2	64,32,512	Deconv IN ReLU	k = 3, s = 2, p = 1 c = 512 –
G^2	64,32,512 × 2	128,64,128	Deconv IN ReLU	k = 3, s = 2, p = 1 c = 128 –
G^1	128,64,128 × 2	256,128,32	Deconv IN ReLU	k = 3, s = 2, p = 1 c = 32 –
G^0	256,128,32	256,128,3	RefPad Conv Tanh	p = 3 k = 7, s = 1, p = 0 –

4.3. Analysis of V-bot

Table 2 shows the comparison between our baseline video model V-BoT and other SOTA methods on Mars, iLIDS-VID, and PRID2011. Although V-BoT is only a concise baseline model, it achieves quite comparable performance with other SOTA methods who claim sophisticated aggregation methods [10,11,17,19–21]. The mAP of V-BoT on Mars surpasses nearly all the methods except BiCnet-TKS [21] (only 0.1% mAP lower). Notably, BiCnet-TKS contains two branches, one takes original resolution frames and the other one takes down-sampled frames as input, and each branch contains multiple parallel and diverse attention modules, whereas V-BoT only uses temporal average pooling. The surprisingly good performance of V-BoT makes us rethink the necessity of designing sophisticated and complicated aggregation models, which are currently the mainstream in the community.

4.4. Analysis of SFGN

4.4.1. Effectiveness of SFGN

We design SFGN to encode the sequence-level features into super frames and decode them when retrieving. Each video sequence is fully represented by one generated super frame. SFGN contains a video model and an image model, and we represent them as SFGN_V and SFGN_I, separately. The performance of SFGN_I is the final performance of super frames. To prove the effectiveness of SFGN, we need to answer **three** questions.

First, whether the performance of the super frames is better than that of directly picking or pixel-level averaging? These two straightforward methods are the lower bounds, and the performance of our super frames should be better than theirs at least.

For directly picking, as in [22], we choose the first frame of the sequence as the representation of the sequence, which is represented as P_1, where "P" is the abbreviation of "pick". For pixel-level averaging, we average the first 4 frames of each sequence, which is represented as $\text{P}^{\text{avg}}_{1:4}$. For fair comparison with super frames, we use V-BoT (T = 1) to train the lower bounds.

Their results on Mars are illustrated in Table 3. By comparing $\text{P}^{\text{avg}}_{1:4}$ with P_1, we find that averaging the frames at pixel-level directly will reduce the performance significantly. With the pixel-level averaging, the mAP of $\text{P}^{\text{avg}}_{1:4}$ decreases by 17.8% compared with P_1. As illustrated in Fig. 3, the pixel-level averaged image is blurry and of low quality. Thus, the unsatisfactory performance is reasonable. By comparing SFGN_I with P_1, we find that our SFGN_I outperforms P_1 by a large margin (25.3% in mAP and 20.3% in Rank-1), which proves that the generated super frames are much powerful than the directly picked frames.

Second, how close the performance of super frames is to that of the sequence-level features? The performance of the sequence-level features (i.e., the performance without considering the super frame generating) is the theoretical upper bound for that of super frames. Thus, the closer their performances are, the more effective SFGN is. Considering that V-BoT is used as the video model in SFGN, the performance of the independently trained V-BoT is reported as the upper bound.

By comparing SFGN_I with V-BoT in Tables 2 and 3, we find that the performance of the decoded sequence-level features from super frames (SFGN_I) is on par with that of V-BoT, which means that the encoding–decoding mechanism of our proposed SFGN suffers negligible information loss. The mAP of SFGN_I is 0.3% lower than that of V-BoT, but the Rank-1 is 1.9% higher than that of SFGN_I. We believe this is reasonable because V-BoT is only the theoretical upper bound instead of the strict one, and the performance of deep models often fluctuates within a certain range.

Third, whether the super frame training process will decrease the performance of the co-trained video model?

By comparing SFGN_V with V-BoT in Table 3, we find that the co-trained video model achieves similar performance as the independently trained video model, which proves that the super frame training process will **not** reduce the performance of the co-trained video model.

In summary, the proposed SFGN can generate powerful super frames whose performance significantly outperforms that of directly picking or pixel-level averaging, and is on par with those of the sequence-level features. Besides, the super frame training process will not reduce the performance of the co-trained video model. Thus, the proposed SFGN is powerful and effective.

4.4.2. Effectiveness of SFSG

SFSG is mainly composed of skip-connection and specific-frame-feature fusion. Skip-connection is proposed to compensate for the fine-grained information loss which is caused by the down-sampling operation, and specific-frame-feature fusion is

Table 2
Comparison of our baseline video model (V-BoT) and super frame ($SFGN_I$) with other SOTA methods on Mars, iLIDS-VID and PRID2011. Notably, V-BoT represents the performance of independently trained sequence-level features, which is the theoretical upper bound of $SFGN_I$. The performance of $SFGN_I$ is on par with V-BoT, which shows that the designed SFGN can successfully generate super frames of high representation ability. Also, although V-BoT only utilizes the simple temporal average pooling (TAP) as the aggregation model and does not intend to boost the performance, it achieves quite competitive performance compared with other SOTA methods that utilize elaborately designed aggregation methods. This makes us rethink the necessity of designing sophisticated aggregation models, which is the main stream in previous works.

		Mars				iLIDS-VID			PRID2011		
Method	Aggregation	mAP	R-1	R-5	R-20	R-1	R-5	R-20	R-1	R-5	R-20
QAN (CVPR 2017) [27]	Attention	–	–	–	-	68.0	86.8	97.4	90.3	98.2	**100**
ID-aware (arXiv 2019) [19]	Attention	71.7	83.3	–	-	81.9	–	-	93.7	–	–
RRU+STIM (AAAI 2019) [14]	RNN	72.7	84.4	93.2	96.3	84.3	96.8	99.5	92.7	98.8	99.8
M3D (AAAI 2019) [13]	3DCNN	74.1	84.4	93.8	97.7	74.0	94.3	-	94.4	**100.0**	–
GLTR (ICCV 2019) [26]	Attention	78.5	87.0	95.8	98.2	86.0	98.0	-	95.5	**100.0**	–
STA (AAAI 2019) [18]	Attention	80.8	86.3	95.7	98.1	–	–	-	–	–	–
STMN (ICCV 2021) [20]	Attention	84.5	**90.5**	–	-	–	–	-	-	-	–
MSF (arXiv 2019) [11]	Attention	85.2	87.1	–	-	87.7	–	-	95.8	–	–
MG-RAFA (CVPR 2020) [10]	Attention	85.9	88.8	**97.0**	**98.5**	88.6	98.0	99.7	95.9	99.7	**100**
BiCnet-TKS (CVPR 2021) [21]	Attention	**86.0**	90.2	–	-	–	–	-	–	–	–
V-BoT (our concise video model)	TAP	85.9	88.0	96.7	98.4	**91.5**	**98.1**	**99.9**	**96.0**	98.8	**100**
$SFGN_I$ (our super frame)	TAP	85.6	**89.9**	96.4	98.2	90.4	**98.7**	**99.9**	95.5	98.9	99.9

Table 3
Comparison of SFGN with baselines on Mars. "Strategy" refers to the approach for obtaining the frames for inference. "Number" represents for the number of frames used per sequence for inference. The performance of our super frames ($SFGN_I$) significantly outperforms that of directly picking (P_1) and is close to that of sequence-level features (V-BoT).

Method	Strategy	Number	mAP	R-1	R-5	R-10	R-20
$P^{avg}_{1:4}$	Pick First 4 + Pixel-level Average	1	42.5	55.3	76.0	82.1	86.4
P_1 (Lower Bound)	Pick First	1	60.3	69.6	87.6	91.0	93.0
$SFGN_I$	Super Frame	1	85.6	**89.9**	96.4	**97.5**	98.2
$SFGN_V$	Dense	all	85.8	89.0	96.3	97.4	98.0
V-BoT (Upper Bound)	Dense	all	**85.9**	88.0	**96.7**	97.4	**98.4**

Table 4
Ablation study of FRL on Mars. "w/o" represents for "without". Without FRL, the performance of super frames is lower and the information loss between the image model and video model is larger.

Model	mAP	R-1	R-5	R-10	R-20
$SFGN_I$ w/o FRL	84.1	**88.7**	**96.6**	**97.6**	**98.2**
$SFGN_V$ w/o FRL	**84.6**	88.0	95.5	96.8	97.6
$SFGN_I$	85.6	**89.9**	**96.4**	**97.5**	**98.2**
$SFGN_V$	**85.8**	89.0	96.3	97.4	98.0

proposed to solve the problem of average-looking due to the aggregation operation. To illustrate the impact of each part qualitatively, we visualize the super frames in Fig. 3. There exist two kinds of possible reconstruction labels for SFSG, namely, the pixel-level averaged one, and the directly picked one (we use the middle frame by default). We represent them as F_{avg_p} and F_{mid}, separately. Although F_{mid} is taken as the reconstruction label in our final version, we illustrate both of their results to help readers understand the effectiveness of SFSG better.

By observing Fig. 3, we can obtain the conclusions as follows:

- By comparing Avg+Skip with Avg, and Mid+Skip with Mid, we find that the skip-connection can largely make up for the fine-grained information loss caused by the down-sampling operation. This idea is similar to the multiple knowledge representations [52] which can enhance the visual quality.
- By observing Avg+Skip and Mid+Skip, we find that the images directly reconstructed from the sequence-level features are intrinsically average-looking. The reconstructed images still tend to be average-looking even when a clean frame is used as the reconstruction label (Mid+Skip). This motivates us to change the paradigm from "directly generating" to "encoding" and propose the specific-frame-feature fusion.
- By comparing Mid+Skip with Mid+Skip+Fusion, we find that after adding the specific-frame-feature fusion, the high-quality super frame is achieved, thereby proving the importance and effectiveness of the specific-frame-feature fusion.

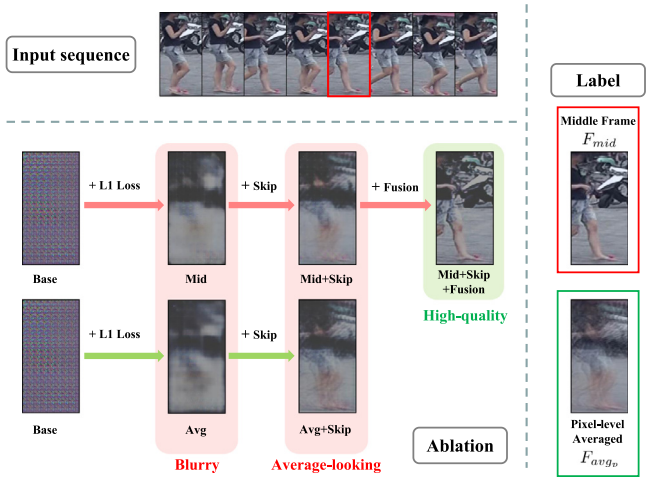
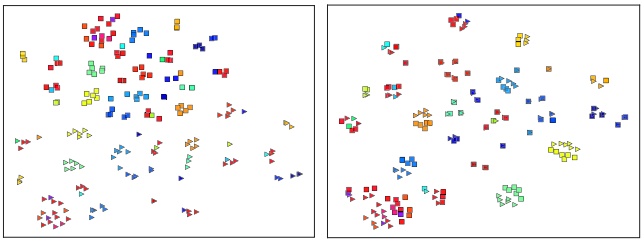


Fig. 3. The ablation of SFSG. The frames highlighted by red and green boxes are the reconstruction labels. The first row is the ablation when taking the middle frame F_{mid} as the reconstruction label, and the second row is the ablation when pixel-level averaged image F_{avg_p} is the reconstruction label. "Mid", "Avg", "Skip", and "Fusion" denote the middle frame label, pixel-level averaged label, skip-connection, and specific-frame-feature fusion, respectively. Mid and Avg generate blurry images, while Mid+Skip and Avg+Skip generate average-looking images. Only SFSG (Mid+Skip+Fusion) can generate a high-quality image. We randomly pick 8 frames from the original sequence for illustration since the entire sequence is too long to illustrate.

More examples of the generated super frames are illustrated in Fig. 4. Obviously, the generated super frames are of high visual quality. We believe that they are clear enough for the users (e.g., police) to verify the retrieval results.

Fig. 4. The examples of generated super frames from dense strategy. Middle frame from a sequence is assigned as the reconstruction label. We can easily find that the generated super frames are of high visual quality.



(a) SFGN w/o FRL (b) SFGN

Fig. 5. The visualization of feature distributions of SFGN w/o FRL and SFGN on the test set of Mars by t-SNE. The features of the first 100 sequences are illustrated. Sequences are sampled by dense strategy. Different colors represent for different identities. Squares represent for the sequence-level features from video model (f_{aft_V}). Triangles represent for the decoded sequence-level features from super frames (f_{aft_I}). Without FRL, the decoded features (triangles) are away from the original sequence-level features (squares), as illustrated in (a). After adding FRL, these two feature distributions are aligned. Best viewed in color with zooming in.

Table 5
Comparison between different strategies for selecting the reconstruction labels in SFGN on Mars.

Strategy	mAP	R-1	R-5	R-20
Random	**85.7**	**89.9**	96.3	**98.2**
Feature center	85.6	89.8	**96.4**	98.1
Middle	85.6	**89.9**	**96.4**	**98.2**

4.4.3. Effectiveness of FRL

To fully decode the sequence-level features, we propose FRL to minimize the L_1 distance between f_{aft_I} and f_{aft_V}. We compare the performance of SFGN with and without FRL (SFGN w/o FRL) in Table 4 to prove the effectiveness of FRL. For each model, we report the performance of the image model ($SFGN_I$, $SFGN_I$ w/o FRL) together with the video model ($SFGN_V$, $SFGN_V$ w/o FRL). The results show that the proposed FRL can boost the performance of super frames and reduce the information loss between the video model and the image model.

To help readers better understand the effectiveness of FRL, we visualize the feature distributions of SFGN and SFGN w/o FRL in Fig. 5. It shows that without FRL, the decoded sequence-level features (f_{aft_I}) are away from the encoded sequence-level features (f_{aft_V}). After adding FRL, these two feature spaces are drawn closer.

4.5. Further analysis

4.5.1. Analysis of the specific frame selection

We name the frame that is used as the reconstruction label as "the specific frame" because the performance of the decoded sequence-level features is irrelevant to the quality of the reconstruction labels. We can specify any frame as the reconstruction label as long as its clean frame-level feature is fused with the sequence-level feature. As proof, we conduct experiments of three different assignment strategies, namely random, feature center, and middle. Random strategy means the reconstruction labels are arbitrarily selected among the sequences, leading to unstable quality of the selected frames (i.e., the frame with occlusion or false detection may be picked). Feature center strategy means that the frame closest to the feature center of the whole sequence is used as the reconstruction label. In this condition, the chosen reconstruction label usually has a high quality. The middle strategy directly uses the middle frames of each sequence as the reconstruction labels. The results in Table 5 show that the performances of these strategies have negligible differences, which means the performance of the decoded features is not related to the visual quality of the assigned specific frames. Moreover, it proves that the super frame visual reconstruction progress is well decoupled with the feature encoding–decoding process, which means SFGN is effective.

Table 6
Comparison between the knowledge propagation version (KP) and the encoding–decoding version ($SFGN_I$) on Mars. The suffix "-L_1" and "-L_2" represent for L_1 distance and L_2 distance, separately.

Method	mAP	R-1	R-5	R-20
KP_{-L1}	76.5	82.0	92.8	96.3
KP_{-L2}	76.0	81.9	92.6	96.0
$SFGN_{I-L1}$	**85.6**	**89.9**	**96.4**	98.2
$SFGN_{I-L2}$	85.4	89.1	**96.4**	**98.3**

4.5.2. Comparison with knowledge propagation

The usage of FRL in SFGN seems similar to that of knowledge propagation in [22], which proposes minimizing the L_2 distance between the frame-level features and the sequence-level features. However, we are fundamentally different. First, they target on solving the I2V problem (i.e., the query is composed of images, the gallery is composed of video sequences, and they ignore the problem of storage overhead); thus, the input of their image model is the original image frames, whereas ours are the generated super frames. Second, they intend to leverage the constrain of L_2 distance to propagate the temporal information in sequence-level features to image-level features, whereas our purpose is to decode the sequence-level features in super frames.

To prove that the effectiveness of super frames is brought by the designed encoding–decoding mechanism instead of knowledge propagation, we replace the super frames with the original middle frames during training. By doing so, FRL can be considered as the temporal knowledge propagation loss. We name the knowledge propagation version as KP, and we compare it with the original SFGN under both L_1 and L_2 distance in Table 6. The results show that L_1 and L_2 distance exhibit similar performance, and SFGN outperforms KP by a large margin (9.1% mAP and 9.4% mAP for L_1 and L_2 distance, respectively). This finding proves that the strong representation ability of our generated super frames is attributed to our elaborately designed encoding–decoding mechanism instead of the knowledge propagation.

Table 7
Comparison between pixel-level fusion (PF) and feature-level fusion (FF) on Mars. The number in between "()" is the performance of the co-trained video model. V-BoT-Mid and V-BoT-Dense represent for using middle frames and dense strategy for inference, separately.

Method	mAP	R-1	R-5	R-20
V-BoT-Mid	77.0	83.6	93.9	96.7
PF	80.9 (80.9)	86.5 (85.3)	95.8 (94.5)	**98.2** (97.3)
FF (ours)	**85.6** (**85.8**)	**89.9** (**89.0**)	**96.4** (**96.3**)	**98.2** (**98.0**)
V-BoT-Dense	85.9	88.0	96.7	98.4

Table 8
The FLOPs (G) and parameters (M) for each part of SFGN. ResNet-50 is used as the backbone. In SFGN, the video model and SFSG can be used offline for generating super frames, while only the efficient image model is used online. We test with the first sequence (40 frames) in the test set of Mars using dense strategy.

Method	FLOPs	Param.
Video Model (V-BoT)	65.24	23.51
SFSG	24.22	67.32
Image Model	1.63	23.51

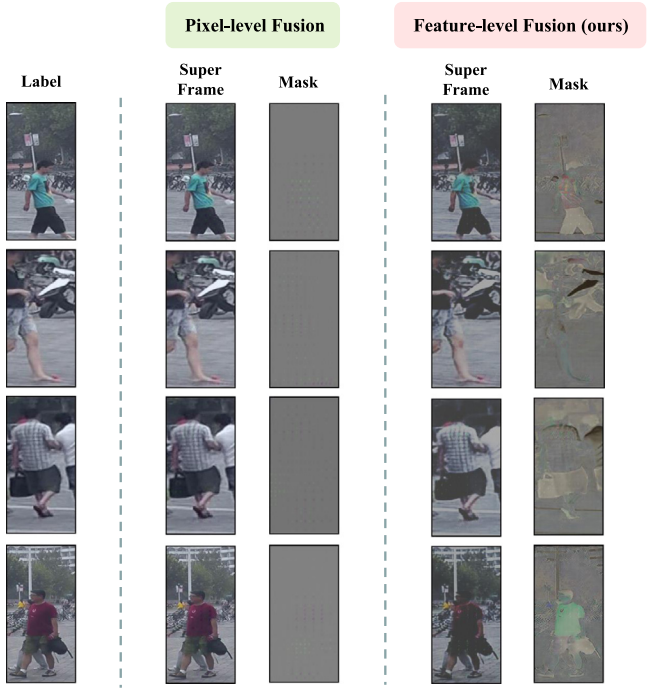


Fig. 6. The comparison of super frames and masks between pixel-level fusion and feature-level fusion. We randomly pick 4 examples from the test set of Mars for illustration. Note that the masks in the feature-level fusion column are manually obtained by the **subtraction** between super frames and corresponding labels for comparison with the generated masks in pixel-level fusion. The observable details on the generated masks of pixel-level fusion are teeny-weeny. Best viewed in color with zooming in.

4.5.3. Feature-level fusion vs. Pixel-level fusion

Our proposed specific-frame-feature fusion in SFSG belongs to the feature-level fusion (FF). However, there seems to be a more straightforward way to fuse the information. Specifically, we can directly generate a mask image from the sequence-level feature, and then add it to the middle frame to get the super frame, which is then constrained by the L_1 reconstruction loss with the middle frame as the label. We name such paradigm as pixel-level fusion (PF). The comparison between PF and FF is illustrated in Table 7. By observing the table, we summarize the conclusions as follows: (1) The performance of FF is significantly better than that of PF, e.g., FF can bring about 8.6% mAP increase compared with using original middle frames (V-BoT-Mid), while PF can only boost the performance by 3.9% in mAP. (2) PF will largely decrease the performance of the co-trained video model, while FF will not. For instance, the mAP of the co-trained video model in PF is 5.0% smaller than that of the independently trained sequence-level features (V-BoT-Dense), while FF only decreases the mAP by 0.1%.

Also, we illustrate the super frames and masks of PF and FF in Fig. 6. Obviously, PF can generate masks with limited information, while FF can achieve a good balance between the high performance and the high visual quality. In summary, FF is a better choice than PF, which is reasonable since FF can aggregate the features better by the nonlinear operation, while the PF is merely an additive operation.

4.5.4. Rethinking the encoding-decoding mechanism

Why do we term the SFGN as an encoding–decoding process? Revisiting the design of SFGN, SFSG is designed to be a feature encoder and the image model is designed to be a feature decoder.

Encoding process. In SFSG, we fuse the sequence-level features with the middle frame features for up-sampling. The middle frame features part makes reconstructing clean super frames possible, and the sequence-level features part makes decoding sequence-level features from super frames possible.

Decoding process. During the training process, we use L1 reconstruction loss to force the super frames to be clean, which leverages the information from the middle frame features part. We also use FRL to force the extracted features from super frames to be close to the original sequence-level features, which leverages the information from the sequence-level features part. Therefore, after the training, the SFSG can generate relatively clean super frames, and the image model achieves the ability of decoding the sequence-level features from the super frames.

Also, two important observations that can support the encoding–decoding hypothesis were made: (1) According to Table 5 and Section 4.5.1, the performance of the super frames is not so related to what the super frames look like, suggesting that it is more reasonable to take the image model as a feature decoder instead of a normal feature extractor, which merely extracts features based on the appearance of images. (2) The performance of super frames is similar to that of the sequence-level features (Table 3), which shows that the output features of the image model share similar performance with sequence-level features.

4.5.5. Why not directly use middle-frame features

In SFGN, we use the middle frames as the reconstruction labels by default. Thus, the super frames are **visually similar** to the middle frames. Then, why do we not directly use the middle-frame features for retrieval? In fact, as illustrated in Table 7, directly using the middle-frame features for retrieval achieves a performance that is significantly lower than that of sequence-level features. Meanwhile, the performance of our super frames (see SFGN_I in Table 2) is similar to that of sequence-level features (see V-BoT in Table 2). As analyzed in Section 4.5.4, SFGN is an encoding–decoding process. Therefore, it will be a better way to understand by considering the image model as the corresponding feature decoder for the encoder (SFSG) instead of a normal feature extractor.

4.5.6. Efficiency analysis

To analyze the efficiency quantitatively, we report the FLOPs (G) and the number of parameters (M) of SFGN in Table 8. For typical methods, if the video sequences are stored in the gallery and the video model is used online for extracting video-level features, the FLOPs will be 65.24 G. While in SFGN, the online image model is much more efficient (1.63 G in FLOPs), and the video model and SFSG can be used offline.

Table 9
The performance comparison between super frames ($SFGN_I$) and the independently trained video model (V-BoT) when combined with PSTA [53].

Method	mAP	R-1	R-5	R-20
$SFGN_I$ (PSTA)	84.4	**89.1**	**96.4**	**98.4**
V-BoT (PSTA)	**86.1**	89.0	96.0	98.0

Fig. 7. Visualization of super frames when using different aggregation methods in SFGN. PSTA [53] is a SOTA sophisticated aggregation method, and TAP represents for temporal average pooling, which is used by default in our work. We randomly pick 4 examples from the test set of Mars for illustration. Obviously, SFGN can still generate high-quality super frames when combined with PSTA, thereby proving the flexibility of SFGN. Best viewed in color with zooming in.

4.5.7. *Flexibility analysis*

To prove the flexibility of SFGN, we replace the temporal average pooling module in V-BoT with PSTA [53] from ICCV 2021, which claims sophisticated aggregation module named STAM. We name the new model as V-BoT (PSTA) and report the performance of the super frames ($SFGN_I$ (PSTA)) together with the independently trained V-BoT (PSTA) in Table 9. We also illustrate the generated super frames in Fig. 7. In general, super frames still achieve comparable performance with the upper bound and show high visual quality, hence proving the flexibility of our SFGN.

5. Conclusion

In this paper, we propose that the high storage overhead stems the applications of V-Re-ID. To relieve this issue, we proposed the SFGN, which can generate high-quality super frames encoded with the information of the video sequences. Storage overhead is largely reduced, whereas performance is reserved with only one super frame stored per sequence. Extensive experiments have been conducted to prove the effectiveness of our proposed methods. In the future, we will further explore the following aspects: (i) **Improved visual quality.** Currently, we only tried the simple L_1 reconstruction loss. We will explore more reconstruction losses (e.g., perceptual loss [54]) or try to import the adversarial model [28], which may further improve the visual quality of super frames. (ii) **Additional datasets.** We will study the applications of SFGN on more datasets, such as vehicle Re-ID datasets VeRi-776 [55] and VehicleNet [56], in the future. (iii) **More discriminative learning objectives.** In this work, we use the ID loss [4] and triplet loss [42] for discriminative training by default. However, there exist many other objectives which have been employed in Re-ID works. For example, circle loss [57] fuses the ID loss [4] and triplet loss [42], while some works combine the ID loss [4] with the contrastive loss [58]. Some face recognition losses can also be used in Re-ID problems, such as sphere loss [59]. Recently, the metric-based losses, such as lifted loss [60] and instance loss [61], also show further performance boost. Our proposed SFGN is theoretically compatible to all of them, and we leave the study of different optimization objectives to the future work.

CRediT authorship contribution statement

Xiao Pan: Conceptualization, Methodology, Writing – original draft, Investigation. **Hao Luo:** Conceptualization, Validation. **Wei Jiang:** Supervision, Funding acquisition. **Jianming Zhang:** Supervision, Funding acquisition. **Jianyang Gu:** Validation, Visualization. **Peike Li:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62173302) and the Autonomous Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2021A05).

References

[1] F.M. Khan, F. Brémond, Person re-identification for real-world surveillance systems, 2016, arXiv preprint arXiv:1607.05975.
[2] S. Roth, B. Schiele, M. Andriluka, People-tracking-by-detection and people-detection-by-tracking, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Los Alamitos, CA, USA, 2008, pp. 1–8, http://dx.doi.org/10.1109/CVPR.2008.4587583.
[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, 2018 pp. 480–496, http://dx.doi.org/10.1007/978-3-030-01225-0_30.
[4] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM Trans. Multimed. Comput. Commun. Appl. 14 (1) (2017) 1–20, http://dx.doi.org/10.1145/3159171.
[5] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, http://dx.doi.org/10.1109/CVPRW.2019.00190.
[6] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimed. (2019) 2597–2609, http://dx.doi.org/10.1109/TMM.2019.2958756.
[7] Y. Ding, H. Fan, M. Xu, Y. Yang, Adaptive exploration for unsupervised person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. 16 (1) (2020) 1–19, http://dx.doi.org/10.1145/3369393.
[8] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, Y. Yang, Progressive learning for person re-identification with one example, IEEE Trans. Image Process. 28 (6) (2019) 2872–2881, http://dx.doi.org/10.1109/TIP.2019.2891895.
[9] Y. Lin, Y. Wu, C. Yan, M. Xu, Y. Yang, Unsupervised person re-identification via cross-camera similarity exploration, IEEE Trans. Image Process. 29 (2020) 5481–5490, http://dx.doi.org/10.1109/TIP.2020.2982826.
[10] Z. Zhang, C. Lan, W. Zeng, Z. Chen, 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, 2020, pp. 10407–10416, http://dx.doi.org/10.1109/CVPR42600.2020.01042.
[11] X. Jiang, Y. Gong, X. Guo, Q. Yang, F. Huang, W. Zheng, F. Zheng, X. Sun, Rethinking temporal fusion for video-based person re-identification on semantic and time aspect, 2019, arXiv preprint arXiv:1911.12512.
[12] J. Gao, R. Nevatia, Revisiting temporal modeling for video-based person reid, 2018, arXiv preprint arXiv:1805.02104.

[13] J. Li, S. Zhang, T. Huang, Multi-scale 3D convolution network for video based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8618–8625, http://dx.doi.org/10.1609/aaai.v33i01.33018618.

[14] Y. Liu, Z. Yuan, W. Zhou, H. Li, Spatial and temporal mutual promotion for video-based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8786–8793, http://dx.doi.org/10.1609/aaai.v33i01.33018786.

[15] W. Song, Y. Wu, J. Zheng, C. Chen, F. Liu, Extended global–local representation learning for video person re-identification, IEEE Access 7 (2019) 122684–122696, http://dx.doi.org/10.1109/ACCESS.2019.2937974.

[16] L. Zhang, Z. Shi, J.T. Zhou, M.-M. Cheng, Y. Liu, J.-W. Bian, Z. Zeng, C. Shen, Ordered or orderless: A revisit for video based person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. (2020) http://dx.doi.org/10.1109/TPAMI.2020.2976969.

[17] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018 pp. 369–378, http://dx.doi.org/10.1109/CVPR.2018.00046.

[18] Y. Fu, X. Wang, Y. Wei, T. Huang, STA: Spatial-temporal attention for large-scale video-based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8287–8294, http://dx.doi.org/10.1609/aaai.v33i01.33018287.

[19] X. Wang, E. Kodirov, Y. Hua, N.M. Robertson, ID-aware quality for set-based person re-identification, 2019, arXiv preprint arXiv:1911.09143.

[20] C. Eom, G. Lee, J. Lee, B. Ham, Video-based person re-identification with spatial and temporal memory networks, 2021, arXiv preprint arXiv:2108.09039.

[21] R. Hou, H. Chang, B. Ma, R. Huang, S. Shan, BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2014–2023, http://dx.doi.org/10.1109/CVPR46437.2021.00205.

[22] X. Gu, B. Ma, H. Chang, S. Shan, X. Chen, Temporal knowledge propagation for image-to-video person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9647–9656, http://dx.doi.org/10.1109/ICCV.2019.00974.

[23] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision, Springer, 2014 pp. 688–703, http://dx.doi.org/10.1007/978-3-319-10593-2_45.

[24] W. Zhang, S. Hu, K. Liu, Z. Zha, Learning compact appearance representation for video-based person re-identification, IEEE Trans. Circuits Syst. Video Technol. 29 (8) (2018) 2442–2452, http://dx.doi.org/10.1109/TCSVT.2018.2865749.

[25] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015, http://dx.doi.org/10.1109/CVPR.2015.7298832.

[26] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3958–3967, http://dx.doi.org/10.1109/ICCV.2019.00406.

[27] Y. Liu, J. Yan, W. Ouyang, Quality aware network for set to set recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5790–5799, http://dx.doi.org/10.1109/CVPR.2017.499.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680, http://dx.doi.org/10.1145/3422622.

[29] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762, http://dx.doi.org/10.1109/ICCV.2017.405.

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

[31] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, VRSTC: Occlusion-free video person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7183–7192, http://dx.doi.org/10.1109/CVPR.2019.00735.

[32] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 650–667, http://dx.doi.org/10.1007/978-3-030-01240-3_40.

[33] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, in: Advances in Neural Information Processing Systems, 2017, pp. 406–416, http://dx.doi.org/10.5555/3294771.3294810.

[34] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferrable person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4099–4108, http://dx.doi.org/10.1109/CVPR.2018.00431.

[35] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al., FD-GAN: Pose-guided feature distilling gan for robust person re-identification, in: Advances in Neural Information Processing Systems, 2018, pp. 1222–1233, http://dx.doi.org/10.5555/3326943.3327056.

[36] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019 pp. 2138–2147, http://dx.doi.org/10.1109/CVPR.2019.00224.

[37] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88, http://dx.doi.org/10.1109/CVPR.2018.00016.

[38] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 994–1003, http://dx.doi.org/10.1109/CVPR.2018.00110.

[39] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232, http://dx.doi.org/10.1109/ICCV.2017.244.

[40] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 13001–13008, http://dx.doi.org/10.1609/aaai.v34i07.7000.

[41] G. Chen, Y. Rao, J. Lu, J. Zhou, Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? in: European Conference on Computer Vision, Springer, 2020, pp. 660–676, http://dx.doi.org/10.1007/978-3-030-58598-3_39.

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

[43] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015 pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[44] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134, http://dx.doi.org/10.1109/CVPR.2017.632.

[45] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: A video benchmark for large-scale person re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 868–884, http://dx.doi.org/10.1007/978-3-319-46466-4_52.

[46] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis, vol. 6688, Springer, 2011, pp. 91–102, http://dx.doi.org/10.1007/978-3-642-21227-7_9.

[47] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645, http://dx.doi.org/10.1109/TPAMI.2009.167.

[48] A. Dehghan, S. Modiri Assari, M. Shah, 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, 2015, pp. 4091–4099, http://dx.doi.org/10.1109/CVPR.2015.7299036.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.

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

[52] Y. Yang, Y. Zhuang, Y. Pan, Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies, Front. Inf. Technol. Electron. Eng. 22 (12) (2021) 1551–1558.

[53] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, D. Wang, Pyramid spatial-temporal aggregation for video-based person re-identification, in: ICCV, 2021.

[54] S. Wu, C. Rupprecht, A. Vedaldi, Unsupervised learning of probably symmetric deformable 3D objects from images in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1–10.

[55] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: 2016 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2016, pp. 1–6.

[56] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, T. Mei, VehicleNet: Learning robust visual representation for vehicle re-identification, IEEE Trans. Multimed. 23 (2020) 2683–2693.
[57] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6398–6407.
[58] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, CVPR'06, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006 pp. 1735–1742.
[59] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 212–220.
[60] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012.
[61] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimed. Comput. Commun. Appl. 16 (2) (2020) 1–23.