Context Sensing Attention Network for Video-based Person Re-identification

KAN WANG, UBTECH Research, UBTECH Robotics, Inc., China and Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
CHANGXING DING, South China University of Technology
JIANXIN PANG, UBTECH Research, UBTECH Robotics, Inc
XIANGMIN XU, School of Future Technology, South China University of Technology

Video-based person re-identification (ReID) is challenging due to the presence of various interferences in video frames. Recent approaches handle this problem using temporal aggregation strategies. In this work, we propose a novel Context Sensing Attention Network (CSA-Net), which improves both the frame feature extraction and temporal aggregation steps. First, we introduce the Context Sensing Channel Attention (CSCA) module, which emphasizes responses from informative channels for each frame. These informative channels are identified with reference not only to each individual frame, but also to the content of the entire sequence. Therefore, CSCA explores both the individuality of each frame and the global context of the sequence. Second, we propose the Contrastive Feature Aggregation (CFA) module, which predicts frame weights for temporal aggregation. Here, the weight for each frame is determined in a contrastive manner: i.e., not only by the quality of each individual frame, but also by the average quality of the other frames in a sequence. Therefore, it effectively promotes the contribution of relatively good frames. Extensive experimental results on four datasets show that CSA-Net consistently achieves state-of-the-art performance.

CCS Concepts: • Computing methodologies → Artificial intelligence; • Artificial intelligence → Computer vision; • Computer vision → Computer vision tasks; • Computer vision tasks → Biometrics;

Additional Key Words and Phrases: Video-based person re-identification, channel attention, feature aggregation

This work was supported by the National Natural Science Foundation of China under Grant 62076101 and U2013601, the Guangdong Provincial Key Laboratory of Human Digital Twin (2022B121201004), and Key-Area Research and Development Program of Guangdong Province, China (2019B010154003), and the Program of Guangdong Provincial Key Laboratory of Robot Localization and Navigation Technology (2020B121202011), and the Natural Science Foundation of China (U21A20487), and Shenzhen Technology Project (JCYJ20180507182610734, KCFZ20201221173411032, Y795001001), and CAS Key Technology Talent Program, and Guangdong Technology Project (No. 2016B010125003).

Authors’ addresses: K. Wang and J. Pang, Floor 13, Building B1, Nanshan iPark, No. 1001 Xueyuan Avenue, Nanshan District, Shenzhen, China, 518055; emails: kan.wang@ubtrobot.com, walton@ubtrobot.com; C. Ding (corresponding author), Room 601, Building 30, South China University of Technology, No. 381 Wushan Road, Tianhe District, Guangzhou, China, 510641; email: chxding@scut.edu.cn; X. Xu, Room 626, Building 30, South China University of Technology, No. 381 Wushan Road, Tianhe District, Guangzhou, China, 510641; email: xmxu@scut.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.
1551-6857/2023/03-ART143 $15.00
https://doi.org/10.1145/3573203

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 4, Article 143. Publication date: March 2023.
1 INTRODUCTION

The goal of video-based person re-identification (ReID) is to identify a person of interest using video sequences captured across disjoint camera views [8, 9, 12, 13, 23, 33, 44, 49–52, 54–56, 61, 64–66]. Compared with individual images, video sequences provide richer cues about pedestrians’ identity; therefore, video-based ReID has become an important topic with the widespread usage of surveillance networks [2, 14, 26, 30–32, 38, 40, 42, 69]. However, as illustrated in Figure 1, it remains a challenging problem due to the presence of interference from other pedestrians, occlusion, and pedestrian detection errors.

A pipeline for video-based ReID typically comprises two sequential steps: frame feature extraction and temporal aggregation. The first of these steps extracts features from each individual frame, and the second one aggregates these features into a video feature. Existing approaches typically focus on the second step [24, 34, 60, 62, 63, 67], and are largely temporal pooling-based. Temporal pooling-based methods predict an attention score for each frame in order to promote the contribution of high-quality frames and suppress that of poor-quality ones. However, these methods usually estimate the score based on the content of each frame alone, often ignoring the temporal context [34, 60]. Subsequent approaches have improved frame features by means of information propagation between frames [24, 62, 63]. However, information propagation tends to occur between semantically similar frames, meaning that poor-quality frame features may be revised by other poor-quality ones. Moreover, the above works may underestimate the importance of the frame feature extraction step.

In this work, we demonstrate the advantages of improving both the frame feature extraction and temporal aggregation steps. Our key observation is that the content in video sequence is stable; therefore the temporal context is consistent across most frames. Accordingly, we propose the Context Sensing Attention Network (CSA-Net), which constructs two novel modules that respectively extract discriminative and robust frame- and video-level representations.

First, we propose the Context Sensing Channel Attention (CSCA) module, which extracts robust frame-level features. Existing channel attention modules, such as the squeeze-and-excitation (SE) module [22], aim at emphasizing responses from informative channels and suppress responses from less useful ones. However, without temporal context, such a module may not correctly infer which channels are informative. It becomes clear only when we consider the temporal context information of the entire video sequence. CSCA handles the above problem by modulating the responses of the hidden layer in the SE module according to the overall content of one video sequence. In this way, macro-visual patterns that are irrelevant to the dominant pedestrian are suppressed. Moreover, the responses of the CSCA output layer are free from direct modulation, which enables the individuality associated with each single frame to be taken into account.

Second, we further propose the Contrastive Feature Aggregation (CFA) module for robust temporal aggregation. More specifically, this module consists of two sequential steps. The first step estimates a frame-to-video similarity for each frame. Afterward, the second step determines the weight of each frame by simultaneously considering its own frame-to-video similarity and the average frame-to-video similarities of the other frames in the same video. Therefore, compared with previous methods [34, 60, 67], the frame weight in CFA is predicted in a contrastive manner,
enabling the importance of comparatively better frames to be more effectively highlighted.

Moreover, the proposed CFA bridges two types of feature aggregation methods, i.e., temporal pooling-based [34, 67] and information propagation-based [24, 62, 63].

Extensive experiments are conducted on four video-based ReID benchmarks, i.e., MARS [69], DukeMTMC-VideoReID [41, 59], iLIDS-VID [57], and LS-VID [24]. Experimental results demonstrate the effectiveness of each component in CSA-Net and show that CSA-Net consistently achieves state-of-the-art performance on these databases.

2 RELATED WORK

Early video-based ReID approaches tend to exploit temporal cues for video feature extraction. For example, McLaughlin et al. [37] utilized optical flow for video feature extraction. Li et al. [25] and Xu et al. [60] adopted 3D CNN and Recurrent Neural Networks respectively to extract spatial-temporal representations. However, these methods usually result in a large model size or high computational cost. Inspired by the success in image-based ReID [20, 53, 71], more recent works typically begin by extracting features from each individual frame, then adopt various strategies to aggregate the frame features into a single video feature [24, 35, 62, 67]. According to the way in which temporal aggregation is performed, existing approaches can be roughly divided into the following two categories.

Temporal Pooling-based Methods An intuitive strategy involves applying temporal average pooling on frame features to obtain the video feature [59, 69]. However, this strategy suffers from the impact of noisy frames. To address this issue, some works attempt to highlight features from high-quality frames and suppress those from low-quality ones by means of different weighting strategies [28, 34, 45, 60, 67, 74]. For example, Liu et al. [34] introduced a quality aware network (QAN) to associate each frame with a quality score. Li et al. [28] assigned weights to image regions, with each weight based on the visibility. Zhang et al. [67] predicted the weight for each frame feature according to its correlations with the averaged feature maps in a sequence. Although these approaches weaken the impact of noisy frames, they typically focus on the temporal aggregation step while underestimating the importance of the frame feature extraction step. Moreover, these approaches usually estimate the frame weight based on the content of each frame alone [28, 34], while tending to ignore the informative temporal cues. By contrast, we argue that the contribution
of each frame should be determined with reference not only to its own quality, but also to the average quality of the other frames in the video sequence.

**Information Propagation-based Methods** Methods in this category first refine each frame feature using the features of the other frames [24, 29, 62, 63]. After the quality of each frame feature is improved, a simple temporal pooling strategy, such as averaging, can be adopted to obtain the video feature. For example, Li et al. [24] designed a temporal self-attention module to capture the long-term pair-wise relations between frames. The frame features are then revised according to the pair-wise relations between them. Besides, Yan et al. [62] constructed multi-granular hypergraphs that model both the short- and long-term dependencies between frame features, thereby providing more diverse information to improve these frame features. However, when the self-attention mechanism is used, poor-quality frame features are more likely to be revised by semantically similar frames, the quality of which may also be poor.

Unlike the works discussed above, CSA-Net improves both the frame feature extraction and temporal aggregation steps; therefore, the extracted video feature is more discriminative and robust. Moreover, we propose a novel temporal aggregation method that not only achieves excellent performance, but also bridges the temporal pooling-based and information propagation-based methods.

3 CONTEXT SENSING ATTENTION NETWORK

3.1 Overview

CSA and CFA can be deployed on various baselines, including IDE [72], PCB [47], and MPN [6]. For the sake of simplicity, we illustrate the two modules on the IDE baseline in Figure 2. In this IDE model, ResNet-50 [17] is adopted as the backbone; here, the last spatial down-sampling operation is removed, following [47], in order to increase the size of the output feature maps.

As illustrated in Figure 2, CSA-Net takes a sampled video sequence \( I = \{ I_1, I_2, \ldots, I_T \} \) as input. The backbone produces the frame-level feature maps \( \{ F_1, F_2, \ldots, F_T \} \). CSA then refines each \( F_t \), as follows:

\[
\hat{F}_t = F_t \otimes c_t.
\]

Here, \( \otimes \) denotes the channel-wise multiplication operation. \( c_t \), which represents the channel weights obtained by CSA for the \( t \)th frame, considers both the individuality of the \( t \)th frame and the overall content of the entire sequence.

Next, the refined feature maps \( \{ \hat{F}_1, \hat{F}_2, \ldots, \hat{F}_T \} \) are fed into a **Global Average Pooling (GAP)** layer and a **Fully Connected (FC)** layer to obtain the frame features \( \{ f_1, f_2, \ldots, f_T \} \). Finally, these frame features are aggregated to form the video feature \( h \) via weighted averaging:

\[
h = \frac{1}{T} \sum_{t=1}^{T} w_t f_t,
\]

where \( w_t \) denotes the weight estimated by the CFA module for the \( t \)th frame. It is determined in a contrastive manner by considering both the quality of each individual frame and the average quality of the other frames in sequence.

During training, both cross-entropy loss and triplet loss [43] are employed to optimize \( h \), as illustrated in Figure 2. The two-loss terms are realized in the same way as in existing works [62, 63, 67]. During testing, \( h \) is employed as the representation of a video sequence. The cosine metric is adopted for performance evaluation.

3.2 Context Sensing Channel Attention

We propose CSCA to highlight responses from informative channels for each frame and suppress those of less useful channels. As illustrated in Figure 1(a), the image content may vary dramatically
Fig. 2. Architecture of CSA-Net. It includes two novel components, i.e., CSCA and CFA. CSCA is attached directly to the output feature maps of one backbone model, e.g., ResNet-50. It produces frame-level channel weights that highlight responses from relevant channels to the dominant pedestrian in the video sequence; in this way, it promotes the quality of frame features. CFA predicts the weights for frame features and aggregates these features into a single video feature via weighted averaging. The frame weight is computed in a contrastive manner, meaning that it is determined by both the quality of each individual frame and the average quality of the other frames in the sequence. Both cross-entropy loss and triplet loss are employed to optimize the video feature.

across frames, especially when interference exists between pedestrians, which makes it difficult to infer informative channels from each individual frame alone. Fortunately, most frames in a video sequence are free from interference, which inspires us to infer informative channels for each image with the help of global temporal context in the video sequence.

Accordingly, CSCA is designed with two key criteria in mind. First, informative channels for each frame should be relevant to the dominant pedestrian in the sequence. Second, informative channels can vary across frames in the same sequence; this is because of changes in pedestrian appearance due to pose and viewpoint variations. Therefore, CSCA aims at considering both the global context of the entire sequence and the individuality of each frame.

The architecture of CSCA is illustrated in Figure 2. First, we feed the feature maps for each frame to a GAP layer, the obtained feature vector is denoted as \( z_t \) for the \( t \)th frame. Second, CSCA adopts a two-branch structure: one for individual frames and the other for the entire sequence. The first branch processes \( z_t \) using a \( 1 \times 1 \) Conv layer, the parameters of which are denoted as \( W^l \). The output of this layer is denoted as \( z^l_t \). The second branch comprises a temporal average pooling (TAP) layer, one \( 1 \times 1 \) Conv layer, and one sigmoid layer. The two Conv layers do not share parameters. The operation of the second branch can be represented as follows:

\[
\bar{z} = \frac{1}{T} \sum_{t=1}^{T} z_t, \\
z^\theta = \sigma(W^\theta \bar{z}),
\]

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 19, No. 4, Article 143. Publication date: March 2023.
where $\sigma$ denotes the sigmoid function. $W^g \in \mathbb{R}^{r_1 \times C}$ denotes the parameters of the Conv layer in the second branch, while $r_1$ and $C$ denote the reduction ratio and the dimension of $z_t$, respectively. Here, $r_1$ is empirically set to 4.

Third, $z^g$ acts as a gating mechanism to modulate elements in $z^l_t$:

$$\hat{z}_t = z^g \odot z^l_t,$$

where $\odot$ represents the element-wise multiplication operation, which highlights the responses of relevant elements in $z^l_t$ to the dominant pedestrian in the sequence and suppresses responses caused by interferences.

Finally, we employ another $1 \times 1$ Conv layer, which is followed by a sigmoid layer, to obtain the final channel attention for the $t$th frame:

$$c_t = \sigma(W_1 \hat{z}_t),$$

where $W_1 \in \mathbb{R}^{C \times r_1}$ denotes the parameters of the Conv layer. $c_t$ is used to refine $F_t$ according to Equation (1).

Compared with the SE module [22], CSCA only introduces one extra $1 \times 1$ Conv layer; therefore, CSCA is still computationally efficient with a compact model structure.

**Discussion** The channel weights obtained by CSCA for each frame are modulated by the content of the entire video sequence. The modulation occurs on $z^l_t$ rather than $c_t$. According to the difference in dimensionality, the elements in $z^l_t$ and $c_t$ can be interpreted as responses to macro- and micro-visual patterns, respectively. Intuitively, it is easier to infer the identity-relevant macro-visual patterns than the micro ones; this is because the latter may represent the individuality of one frame, while the former are usually stable across frames. In the experimentation section, we empirically prove that it is indeed better to impose the modulation on $z^l_t$ rather than $c_t$.

### 3.3 Contrastive Feature Aggregation

As illustrated in Figures 1(c) and (d), the feature quality of some frames is inherently limited due to occlusion and pedestrian detection errors. We accordingly further propose the CFA module to weaken the influence of poor-quality frames by exploring temporal context information in the temporal aggregation step.

Our key observation here is that, in most videos, only a small fraction of frames are of poor quality, the overall content in the video sequence is therefore stable. This indicates that the consistency between each frame and all frames in sequence can reflect the quality of the individual frame. Inspired by this observation, we design the CFA module. In general, this module comprises two components, i.e., consistency measurement and contrastive weight formulation.

#### 3.3.1 Consistency Measurement. This component takes the frame features $\{f_1, f_2, \ldots, f_T\}$ as input and estimates a quality score for each frame. This score, denoted as $s_t$ for the $t$th frame, measures the average similarity between the $t$th frame and each of the frames in sequence. As shown in Figure 2, this component constructs two parallel $1 \times 1$ Conv layers that share parameters. The output of these two Conv layers is used to compute $X \in \mathbb{R}^{T \times T}$ as follows:

$$x_{ij} = \frac{\theta(f_i)^T \theta(f_j)}{\|\theta(f_i)\| \|\theta(f_j)\|},$$

where $x_{ij}$ denotes the element in the $i$th row and $j$th column of $X$. It represents the cosine similarity between the two frame features $f_i$ and $f_j$. Moreover, $\theta(f_i) = W_2 f_i$, where $W_2 \in \mathbb{R}^{d \times d}$ represents the parameters of the Conv layer. $d$ denotes the dimension of frame features, while $r_2$ represents the reduction ratio, which is empirically set to 2.
Finally, the consistency-aware quality score $s_t$ for the $t$th frame can be obtained as follows:

$$s_t = \frac{1}{T} \sum_{i=1}^{T} x_{ti}. \quad (7)$$

3.3.2 Contrastive Weight Formulation. The next step is to aggregate the frame features. As introduced in Section 2, recent information propagation-based methods [24, 62, 63] first refine frame features according to their pair-wise relations, then apply a simple temporal averaging operation on the refined frame features to obtain the video feature. However, frames with close pair-wise relations tend to be of similar quality. Accordingly, we propose to adaptively improve each frame feature with reference to the high-quality ones as follows:

$$\hat{f}_t = s_t f_t + (1 - s_t) \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} s_i f_i, \quad (8)$$

where $\hat{f}_t$ denotes the refined frame feature for the $t$th frame. $\hat{f}_t$ is made up of two parts. The first part, i.e., $s_t f_t$, denotes the contribution of its original feature $f_t$. Here, a larger value of $s_t$ indicates a greater contribution from $f_t$. The second part introduces the contributions from the other $T-1$ frames according to their respective quality scores. Both parts work cooperatively to improve the original frame features.

Finally, the video feature $h$ is obtained by applying the temporal averaging on the improved frame features $\hat{f}_t$:

$$h = \frac{1}{T} \sum_{t=1}^{T} \left\{ s_t f_t + (1 - s_t) \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} s_i f_i \right\}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \left\{ s_t \left( 2 - \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} s_i \right) \right\} f_t. \quad (9)$$

Please refer to Equation (11) for derivation of Equation (9). Accordingly, the final weight for the $t$th frame feature, i.e., $w_t$, can be decoupled into two parts: $s_t$ and $(2 - \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} s_i)$. The first of these represents an absolute weight for $f_t$. The second indicates the average quality score of all the other frames. This part enables $w_t$ to be determined in a contrastive manner: a higher value indicates that the average quality of the other frames is low, and thus that the importance of the $t$th frame should be further emphasized. These two parts are complementary to each other and work collaboratively for temporal aggregation.

**Derivation of Equation (9)** The video feature $h$ is obtained as follows:

$$h = \frac{1}{T} \sum_{t=1}^{T} \left\{ s_t f_t + (1 - s_t) \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} s_i f_i \right\}. \quad (10)$$

Accordingly, the weight $w_t$ for the $t$th frame feature (i.e., $f_t$) can be formulated as follows:

$$w_t = s_t + \sum_{i=1 \ldots T \mid i \neq t} (1 - s_i) \frac{1}{T - 1} s_t$$

$$= s_t + \frac{1}{T - 1} s_t \sum_{i=1 \ldots T \mid i \neq t} (1 - s_i)$$

$$= s_t \left\{ 1 + \frac{1}{T - 1} \sum_{i=1 \ldots T \mid i \neq t} (1 - s_i) \right\}.$$
\[ \begin{align*}
&= s_t \left\{ 1 + \frac{1}{T - 1} \left( (T - 1) - \sum_{i=1}^{T; i \neq t} s_i \right) \right\} \\
&= s_t \left( 2 - \frac{1}{T - 1} \sum_{i=1}^{T; i \neq t} s_i \right). \tag{11}
\end{align*} \]

**Discussion** Equations (8) and (9) bridge the two categories of temporal aggregation methods reviewed in Section 2. They prove that the information propagation-based methods can be equivalent to the temporal pooling-based methods if poor-quality frame features are refined using high-quality ones. Moreover, unlike existing temporal pooling-based methods, CFA determines the frame weight in a contrastive manner; as demonstrated in the experimentation section, this is a more effective approach.

4 EXPERIMENTS

We evaluate our approach on four challenging benchmarks, namely MARS [69], DukeMTMC-VideoReID [41, 59], iLIDS-VID [57], and LS-VID [24], by following their respective evaluation protocols. The cumulative matching characteristic (CMC) and mean Average Precision (mAP) are adopted as evaluation metrics.

**Mars Dataset.** The Mars dataset is a large-scale benchmark for person ReID. It is an extension of the Market-1501 dataset [70] and comprises 17,503 video sequences belonging to 1,261 identities as well as 3,248 distractor sequences. Videos in this dataset were captured by 6 cameras. Pedestrians were detected using Deformable Part Models [10].

**DukeMTMC-VideoReID dataset.** includes 4,832 video sequences associated with 1,812 identities. Videos in this database were captured by 8 cameras. Bounding boxes of pedestrians were manually annotated.

**iLIDS-VID dataset.** The iLIDS-VID dataset consists of 600 video sequences of 300 identities. Two indoor cameras were utilized to capture the pedestrian sequences. Each video sequence contains 23 to 192 frames. This dataset is very challenging because of the large variations in lighting and viewpoints and cluttered backgrounds.

**LS-VID dataset.** includes 14,943 video sequences of 3,772 identities. A camera network consisting of 3 outdoor cameras and 12 indoor cameras were employed to construct this dataset.

4.1 Implementation Details

We implement the proposed CSA-Net based on the PyTorch framework. A standard stochastic gradient descent optimizer with a weight decay of $5 \times 10^{-4}$ and a momentum value [48] of 0.9 is utilized for model optimization. Fine-tuned from the IDE model [17], CSA-Net is trained in an end-to-end fashion for 350 epochs on each of the four benchmarks, with the learning rate initially set to 0.01 and then multiplied by 0.1 every 100 epochs.

All images are resized to 256 × 128 pixels. For data augmentation, we only adopt random erasing [73] with a ratio of 0.5. The margin of the triplet loss is empirically set to 0.25. We sample 4 video sequences for each of the 4 identities to construct a mini-batch; therefore, the batch size is 16. The values of $d$ and $C$ are 512 and 2,048, respectively. During training, we set $T$ to 8; specifically, we sample 8 frames from each sequence by uniformly splitting the sequence into 4 segments and randomly sample 2 frames per segment. During testing, we use all frames of a video to generate the video feature if its length is less than 128; otherwise, 128 frames are sampled following the strategy discussed above.
### 4.2 Ablation Study

We systematically investigate the effectiveness of each key component of CSA-Net on LS-VID, DukeMTMC-VideoReID, and MARS. To ensure comprehensive evaluation, both IDE and MPN [6] are adopted as baselines (for details of MPN, please refer to the appendix). Experimental results are summarized in Table 1.

#### 4.2.1 Effectiveness of CSCA.

In this experiment, we equip the baseline with CSCA only. As shown in Table 1, CSCA yields clear performance improvements across all settings. For example, compared with the IDE baseline, CSCA improves the performance by 3.8% and 4.5% in terms of Rank-1 accuracy and mAP on LS-VID, respectively.

We further support the above quantitative results by visualizing the heat maps for frame-level feature maps produced by the IDE baseline and CSCA-equipped IDE, respectively. As illustrated in the second row of Figure 3, the feature maps produced by IDE show strong responses in the interference regions; by contrast, as the third row of Figure 3 shows, CSCA robustly highlights the body region of the dominant pedestrian in the video sequence. These results demonstrate the effectiveness of CSCA.

#### 4.2.2 Effectiveness of CFA.

In this experiment, we equip the baseline with CFA only. The results listed in Table 1 show that CFA brings consistent performance promotion for both baselines. For example, compared with the IDE baseline, CFA improves the Rank-1 accuracy by 3.4% and mAP by 3.8% on LS-VID.

To further support the above results, we examine the weights learned by CFA with the IDE baseline. As illustrated in Figure 4, the weights produced by CFA are reasonable; for example, CFA assigns lower weights to frames in which parts are missing or occlusions are present. These results justify the effectiveness of CFA.

Finally, we equip the baseline with both CSCA and CFA; this model is referred to as CSA-Net in Table 1. We can observe from the table that CSA-Net consistently outperforms all other models in Table 1. These comparisons validate that CSCA and CFA complement each other.

### 4.3 Further Analysis and Discussions

#### 4.3.1 Comparisons with SENet and FcaNet.

We compare the performance of CSCA with two popular channel attention model, i.e., the SENet [22] and the FcaNet [39]. In the interest of efficient evaluation, only models based on the MPN baseline are evaluated. Two possible designs are compared for each of the two models.

For SENet, the two designs are denoted as “SE-frame” and ”SE-video” respectively in Table 2. “SE-frame” adopts an ordinary SE module to produce channel weights, i.e., \( c_t \), for each respective frame. Informative temporal cues of the sequence are ignored. “SE-video” produces unified channel weights for all frames in the sequence. In more detail, the feature maps (i.e., \( F_t \)) of all \( T \) frames are first temporally averaged, after which channel weights for the averaged feature maps are extracted using an SE module. Subsequently, all \( F_t \) are refined using the obtained channel weights; therefore,
Fig. 3. Visualization of heat maps for feature maps produced by the IDE baseline (images in the second row) and CSCA-equipped IDE (images in the third row). Images in which the IDE baseline fails to focus on the dominate pedestrian are framed in red rectangles.

Table 2. Performance Comparisons with the SE Module and FcaNet

| Method    | LS-VID Rank-1 | mAP | Duke-Video Rank-1 | mAP | MARS Rank-1 | mAP |
|-----------|---------------|-----|-------------------|-----|-------------|-----|
| Baseline  | 82.8          | 70.7| 96.2              | 95.4| 88.5        | 82.5|
| SE-frame  | 83.4          | 71.5| 96.6              | 95.7| 88.7        | 82.6|
| SE-video  | 83.9          | 71.9| 96.7              | 95.9| 89.0        | 83.0|
| Fra-frame | 83.7          | 72.0| 96.7              | 96.0| 88.9        | 82.9|
| Fra-video | 84.4          | 72.3| 96.9              | 96.0| 89.2        | 83.4|
| CSCA      | 84.9          | 72.8| 97.4              | 96.4| 89.7        | 84.1|

“SE-video” ignores the individuality of single frames. To facilitate fair comparison, the structure of the SE modules in both “SE-frame” and “SE-video” are the same as that for CSCA. For FcaNet, the two designs are denoted as “Fca-frame” and “Fca-video”, respectively. The two models are constructed in the same way as that for SENet by simply replacing the SE module with the multispectral channel attention module in FcaNet [39].

Comparison results are tabulated in Table 2. The following observations can be made. First, FcaNet accieves better performance over SE for both designs. This is because more powerful
channel attention module is used in FcaNet. Second, the performance of both “Fra-frame” and “Fra-video” are inferior to that of CSCA. For example, on LS-VID, CSCA outperforms “Fra-frame” by 1.2% and 0.8% in terms of Rank-1 accuracy and mAP, respectively. This is because CSCA not only considers the individuality of each frame, but also the overall content of the entire sequence; as a result, the channel weights produced by CSCA are more reasonable. These results demonstrate the superiority of CSCA.

Finally, we compare the parameter quantity for “SE-frame” and CSCA. When evaluating the parameter quantity, the values of hyper-parameters are set according to the implementation details in Section 4.1. More specifically, the parameter quantity for CSCA and the original SE module [r1] are $3.1M$ and $2.1M$, respectively, indicating that CSCA is still computationally efficient.

4.3.2 CSCA vs. SM. In Table 3, we compare the performance of CSCA with the Spatial Memory (SM) module [7], one method that refines frame-level person representations by suppressing features for distracting scene details. Implement of SM follows [7] and other experimental settings remain unchanged for a fair comparison.

We can observe that CSCA consistently outperforms the SM module on all three benchmarks. This is because SM operates on each frame independently; this means it is not capable of capturing temporal contexts in video sequences. In comparison, CASA not only explores the individuality of

Table 3. Performance Comparisons with the SM Module and one Variant of CSCA

| Method   | LS-VID Rank-1 | LS-VID mAP | Duke-Video Rank-1 | Duke-Video mAP | MARS Rank-1 | MARS mAP |
|----------|---------------|-------------|-------------------|----------------|-------------|----------|
| Baseline | 82.8          | 70.7        | 96.2              | 95.4           | 88.5        | 82.5     |
| CSCA-v   | 84.2          | 72.3        | 97.0              | 96.1           | 89.3        | 83.5     |
| SM       | 83.9          | 72.0        | 96.8              | 96.0           | 89.1        | 83.4     |
| CSCA     | 84.9          | 72.8        | 97.4              | 96.4           | 89.7        | 84.1     |
4.3.3 Comparisons with Variant of CSCA. We compare the performance of CSCA with one possible variant. This variant, denoted as “CSCA-v”, adopts two SE modules to generate frame- and video-level channel weights, respectively; the ways to learn frame- and video-level channel weights are the same as those adopted in “SE-frame” and “SE-video”, respectively. The video-level channel weights are subsequently used to modulate each frame-level channel weights via element-wise multiplication. Accordingly, the essential difference between “CSCA-v” and CSCA lies in the modulation position. To facilitate fair comparison, the other implementation details of “CSCA-v” are kept the same as that in CSCA.

After assessing the results presented in Table 3, we conclude that it is more effective to modulate channel weights in the hidden (first) Conv layer of the SE module. This may be because, compared with the output layer of the SE module, this hidden layer is more compact; its elements can thus be regarded to stand for macro-visual patterns, which are more coherent across frames in a video sequence. These comparisons demonstrate the effectiveness of CSCA.

4.3.4 Comparisons with QAN and Variant of CFA. We compare the performance of CFA with QAN and one possible variant. The variant is denoted as “CFA-v” in Figure 5. QAN computes the weight for each frame based on its own content using an approach adopted in [34]. In brief, it comprises a 1 × 1 Conv layer with an output dimension of 1 and a sigmoid layer for normalization. The input to QAN is the individual feature of each frame. For its part, “CFA-v” adopts the frame-to-video similarity $s_t$ computed in Equation (7) as the weight for the $t$-th frame.

From the comparisons presented in Figure 5, we can make the following observations. First, both QAN and “CFA-v” outperform the baseline, which demonstrates the effectiveness of...
Table 4. Performance Comparisons with the Non-local Module and TM Module

| Method  | LS-VID | Duke-Video | MARS |
|---------|--------|------------|------|
|         | Rank-1 mAP | Rank-1 mAP | Rank-1 mAP |
| Baseline| 82.8 70.7 96.2 95.4 88.5 82.5 |
| Non-local| 83.5 71.3 96.6 95.7 88.9 82.8 |
| TM | 83.9 71.9 96.9 96.1 89.4 83.4 |
| CFA | 84.6 72.5 97.3 96.4 89.6 83.9 |

weighting strategy. Second, “CFA-v” beats QAN. This is because the weight predicted by "CFA-v" is based on temporal cues, meaning that it is easier for "CFA-v" to identify frames that have been contaminated by interference. Third, CFA surpasses "CFA-v". This result validates the superiority of the contrastive weighting strategy, which considers both the quality of an individual frame and the average quality of the other frames in sequence.

4.3.5 CFA vs. Non-local. As illustrated in Equation (8), the proposed CFA module also works in the manner of information propagation. Therefore, in this experiment, we compare CFA with one of the most popular information propagation-based methods, i.e., the non-local module [58]. Implementation of the non-local module follows [24]. The other experimental settings remain unchanged to facilitate a clean comparison.

From the results presented in Table 4, we can observe that CFA consistently outperforms the non-local module on all three benchmarks. This is because the non-local module inherently tends to refine a poor-quality frame feature using these semantically similar frames, indicating that poor-quality frame features may be revised by ones of similar quality. In comparison, CFA adaptively improves poor-quality frame features with reference to the high-quality ones. These experimental results justify the superiority of CFA.

4.3.6 CFA vs. TM. We further compare the performance of CFA with the Temporal Mermory (TM) module [7], which learns a temporal attention to aggregate the frame-level features into a video-level one by focusing more on discriminative frames. Implement of SM follows [7] and other experimental settings remain unchanged for a fair comparison.

It can be seen from the results tabulated in Table 4 that CFA consistently beats the TM module on all three benchmarks. This is because TM determines the importance of each frame in the sequence based on the temporal contexts, while the individuality of this frame is not sufficiently utilized. In comparison, CFA considers both the quality of an individual frame and the overall quality of the other frames in sequence. The above comparisons convincingly demonstrate the superiority of the CFA module.

4.3.7 Impact of the Hyper-parameter T. In this experiment, we evaluate the performance of the proposed CSA-Net at different values of $T$ (namely 4, 6, 8, 12, and 16). All the other experimental settings remain unchanged in order to facilitate a clean comparison.

From the experimental results illustrated in Figure 6, we can make the following observations. First, the performance of CSA-Net tends to be better at higher values of $T$; this is because a longer sequence provides a more stable temporal context. Second, the performance of CSA-Net is robust to the value of $T$ if the value is sufficiently large. In light of these results, we set $T$ as 8 after considering the tradeoff between performance and computational efficiency.
4.4 Comparisons with State-of-the-Art Methods

State-of-the-art approaches to video-based ReID usually extract part features \[62, 63\] or utilize backbones more powerful than ours (AP3D \[15\] adopts a 3D-based ResNet-50 as the backbone, while the plain ResNet-50 is used in CSA-Net). To facilitate fair comparison, we adopt MPN \[6\] that extracts part features as the baseline in this subsection, since its performance is comparable to that obtained by baselines of recent methods \[4, 15\]. Comparisons between CSA-Net and state-of-the-art methods are presented in Table 5. From the table, it can be seen that CSA-Net consistently achieves state-of-the-art performance on each dataset. Specifically, on the DukeMTMC-VideoReID dataset, CSA-Net outperforms one of the most recent methods (i.e., AFA \[4\]) by 0.5% and 1.3% in terms of Rank-1 accuracy and mAP, respectively. On the MARS database, CSA-Net achieves the best Rank-1 accuracy of 90.4%. Besides, CSA-Net also surpasses state-of-the-art approaches on iLIDS-VID by at least 1.4% in terms of Rank-1 accuracy. These comparisons demonstrate the superiority of CSA-Net.

Besides, we extend CSA-Net to learn multi-granularity representation from the spatial perspective. The strategy to extract multi-granularity spatial feature is similar to that adopted in CDPM \[53\]. Experimental results show that the multi-granularity representation further improve the mAP for CSA-Net: for example, the mAP is improved by 1.2% (85.7%–84.5%) on MARS. This indicates that CSA-Net achieves comparable mAP performance compared with SOTA methods. These comparisons firmly demonstrate the progressiveness of the CSA-Net.

Moreover, CSA-Net suppresses two recent part-based methods, i.e., STGCN \[63\] and MGH \[62\], on three datasets for the Rank-1 accuracy. Besides, compared with AP3D \[15\], which adopts 3D convolutions, the model structure of CSA-Net is simpler as all its operations are 2D-based.

Finally, we compare CSA-Net with recent approaches on the LS-VID \[24\] dataset. As LS-VID was released only recently, few works have reported their performance on this dataset. It can be seen from Table 6 that CSA-Net outperforms all comparison methods by significant margins in terms of Rank-1 accuracy. For example, CSA-Net outperforms BiCnet-TKS \[18\], the most recent method, by 0.7%. These experimental results are consistent with those on the first three datasets. In summary, the above comparisons further validate the effectiveness of CSA-Net.

5 CONCLUSION

In this article, we propose a novel model, named CSA-Net, which improves both the frame feature extraction and temporal aggregation steps for robust video-based ReID. CSA-Net incorporates two novel components, i.e., CSCA and \textbf{Contrastive Feature Aggregation (CFA)}. CSCA effectively highlights informative channels for each frame with reference to the content of the entire sequence.
Table 5. Performance Comparisons on MARS [69], DukeMTMC-VideoReID [41, 59], and iLIDS-VID [57]

| Method        | MARS Rank-1 | MARS Rank-5 | MARS Rank-20 | MARS mAP | Duke-Video Rank-1 | Duke-Video Rank-5 | Duke-Video mAP | iLIDS-VID Rank-1 | iLIDS-VID Rank-5 | iLIDS-VID Rank-20 |
|---------------|-------------|-------------|--------------|----------|------------------|------------------|----------------|-----------------|-----------------|------------------|
| Mars [69]     | 68.3        | 82.6        | 89.4         | 49.3     | -                | -                | -              | -               | -               | -                |
| SeeForest [74]| 70.6        | 90.0        | -            | 50.7     | -                | -                | -              | -               | -               | -                |
| RQEN [45]     | 77.8        | 88.8        | 94.3         | 71.1     | -                | -                | -              | -               | -               | -                |
| EUG [59]      | 80.8        | 92.1        | 96.1         | 67.4     | 83.6             | 94.6             | 78.3           | -               | -               | -                |
| CSA [36]      | 83.4        | 93.4        | 97.4         | 83.3     | 89.3             | 98.3             | 88.5           | 86.3            | 97.4            | 99.7             |
| TKP [16]      | 84.0        | 93.7        | 95.7         | 73.3     | 94.0             | -                | 91.7           | -               | -               | -                |
| M3D [25]      | 84.4        | 93.8        | 97.7         | 74.1     | -                | -                | -              | 74.0            | 94.3            | -                |
| COSAM [46]    | 84.9        | 95.5        | 97.9         | 79.9     | 95.4             | 99.3             | 94.1           | 79.6            | 95.3            | -                |
| Snippt [3]    | 86.3        | 94.7        | 98.2         | 76.1     | -                | -                | -              | 85.4            | 96.7            | 99.5             |
| STA [11]      | 86.3        | 95.7        | -            | 80.8     | 96.2             | 99.3             | 94.9           | -               | -               | -                |
| GLTR [24]     | 87.0        | 95.8        | 98.2         | 78.5     | 96.3             | 99.3             | 93.7           | 86.0            | 98.0            | -                |
| Attribute [68]| 87.0        | 95.4        | 98.7         | 78.2     | -                | -                | -              | 86.3            | 97.4            | 99.7             |
| FGRA [5]      | 87.3        | 96.0        | 98.1         | 81.2     | -                | -                | -              | 88.0            | 96.7            | 99.3             |
| VRSTC [21]    | 88.5        | 96.5        | -            | 82.3     | 95.0             | 99.1             | 93.5           | 83.4            | 95.5            | 99.5             |
| MG-RAFA [67]  | 88.8        | 97.0        | 98.5         | 85.9     | -                | -                | -              | 88.6            | 98.0            | 99.7             |
| TACAN [27]    | 89.1        | 96.1        | 98.0         | 84.0     | 96.2             | 99.4             | 95.4           | 88.9            | -               | -                |
| TCLNet [19]   | 89.8        | -           | -            | 85.1     | 96.9             | -                | 96.2           | 86.6            | -               | -                |
| STGCN [63]    | 90.0        | 96.4        | 98.3         | 83.7     | 97.3             | 99.3             | 95.7           | -               | -               | -                |
| MGH [62]      | 90.0        | 96.7        | 98.5         | 85.8     | -                | -                | -              | 85.6            | 97.1            | -                |
| AP3D [15]     | 90.1        | -           | -            | 85.1     | 96.3             | -                | 95.6           | 86.7            | -               | -                |
| AFA [4]       | 90.2        | 96.6        | -            | 82.9     | 97.2             | 99.4             | 95.4           | 88.5            | 96.8            | -                |
| BiCNet-TKS [18]| 90.2       | -           | -            | 86.0     | 96.3             | -                | 96.1           | -               | -               | -                |
| STRF [1]      | 90.3        | -           | -            | 86.1     | 97.4             | -                | 96.4           | 89.3            | -               | -                |
| CSA-Net       | 90.4        | 96.7        | 98.5         | 84.5     | 97.7             | 99.4             | 96.7           | 90.0            | 98.3            | 99.8             |

Table 6. Performance Comparisons on the LS-VID Dataset [24]

| Methods        | Rank-1 | Rank-5 | Rank-20 | mAP  |
|----------------|--------|--------|---------|------|
| STMP [11, 24]  | 56.8   | 76.2   | 87.1    | 39.1 |
| M3D [24, 25]   | 57.7   | 76.1   | 88.2    | 40.1 |
| PCB [47]       | 75.7   | 89.9   | 94.7    | 62.6 |
| GLTR [24]      | 63.1   | 77.2   | 88.4    | 44.3 |
| TCLNet [19]    | 81.5   | -      | -       | 70.3 |
| AFA [4]        | 84.5   | -      | -       | 73.2 |
| BiCNet-TKS [18]| 84.6   | -      | -       | 75.1 |
| **CSA-Net**    | **85.3**| **93.8**| **97.5**| **73.4**|

For its part, CFA predicts the weight of each frame for temporal aggregation; here, the weight is based on the coherence degree between each frame and the entire sequence, and is adaptively determined in a contrastive manner. Experimental results on four benchmarks demonstrate the effectiveness of CSA-Net. In the future, we plan to explore stronger models, e.g., transformer, to model the relations across frames and predict the weight of each frame. Moreover, we will apply our method to other computer vision tasks, e.g., vehicle ReID.

APPENDIX

A MPN-BASED CSA-NET

As illustrated in Figure 2(a), the MPN baseline [6] also adopts the ResNet-50 [17] model as backbone. Based on the feature maps (i.e., $F$ in Figure 2(a)) produced by the backbone, MPN constructs two...
Fig. 7. (a) Architecture of the MPN baseline. Based on the ResNet-50 backbone, MPN constructs two sub-networks for each body part: one MT sub-network that extracts features from $F$ directly, and one AT sub-network that extracts part features from $F_k$ that is cropped from $F$ according to the part’s coarse location. Sub-network structure for the $K$ body parts is the same. In the interests of simplicity, we only illustrate sub-networks for one part in this figure. Red dash lines denote parameter sharing. (b) Architecture of MPN-based CSA-Net. CSA is imposed on $F$. CFA is deployed on $f_k^t$. paralleltasks for the $k$th ($1 \leq k \leq K$) body part, i.e., one main task (MT) and one auxiliary task (AT). The AT is aware of the coarse location of the $k$th body part and regularizes MT to extract part features directly from $F$. Regularization is achieved via parameter sharing between the $1 \times 1$ Convolutional (Conv) layers in both tasks. We do not utilize the feature space alignment strategy in [6].

As illustrated in Figure 7(a), the sub-networks of the two tasks have different inputs during training. Specifically, the input for MT is $F$. The input for AT is the feature maps for the $k$th body part, i.e., $F_k$ in Figure 7(a). $F_k$ is obtained by uniformly dividing $F$ in the height dimension following [47]. For all the other settings, we simply follow the original configuration of MPN [6]. For example, the dimension of both Conv layers is set as 512.

Given a sampled video sequence $I = \{I_1, I_2, \ldots, I_T\}$, the part features produced by the $k$th MT sub-network are denoted as $\{f_1^k, f_2^k, \ldots, f_T^k\}$ and the part features produced by the $k$th AT sub-network are denoted as $\{\hat{f}_1^k, \hat{f}_2^k, \ldots, \hat{f}_T^k\}$. As illustrated in Figure 7(a), the video features $f^k$ and $\hat{f}^k$ for the $k$th body part are generated by averaging the $T$ part-level frame features, respectively.

Similar to [6], we employ the cross-entropy loss to optimize the part-level video features $f^k$ and $\hat{f}^k$. Besides, the $K$ part-level video features from MTs and ATs are respectively concatenated to form the holistic-level video features $h$ and $\hat{h}$, which are optimized using triplet loss [43]. Therefore, the overall objective function during training for a video sequence can be written as follows:

$$L = \sum_{k=1}^{K} L_{id}^k (f^k) + L_{tp} (h) + \sum_{k=1}^{K} \hat{L}_{id}^k (\hat{f}^k) + \hat{L}_{tp} (\hat{h}).$$

(12)
Here $K$ is empirically set as 4 in the experimentation section of the main article. In the testing stage, the ATs are abandoned and only MTs are used for feature extraction. $h$ is employed as the representation of one video sequence.

As illustrated in Figure 7(b), we deploy the proposed CSCA and CFA modules on MPN at similar locations with the IDE baseline. In brief, CSCA is attached to $F$. CFA is attached to the output of the second $1 \times 1$ Conv layer in each MT, i.e., $f^k_t$.

REFERENCES

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, and Ziyan Wu. 2021. Spatio-temporal representation factorization for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 152–162.

[2] Cuiqun Chen, Meng Ye, Meibin Qi, Jingjing Wu, Yimin Liu, and Jianguo Jiang. 2022. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 32, 9 (2022), 6100–6112. DOI: 10.1109/TCST.2022.3157130

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1169–1178.

[4] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. 2020. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In Proceedings of the European Conference on Computer Vision. Springer, 660–676.

[5] Zengqun Chen, Zhiheng Zhou, Junchu Huang, Pengyu Zhang, and Bo Li. 2020. Frame-guided region-aligned representation for video person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence. 10591–10598.

[6] Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. 2022. Multi-task learning with coarse priors for robust part-aware person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3 (2022), 1474–1488.

[7] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. 2021. Video-based person re-identification with spatial and temporal memory networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 12036–12045.

[8] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications 14, 4 (2018), 1–18.

[9] Pengfei Fang, Pan Ji, Lars Pettersson, and Mehrtash Harandi. 2021. Set augmented triplet loss for video person re-identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 464–473.

[10] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 9 (2009), 1627–1645.

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. 2021. 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. 770–778.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 770–778.
[19] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2020. Temporal complementary learning for video person re-identification. In Proceedings of the European Conference on Computer Vision. Springer, 388–405.

[20] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9317–9326.

[21] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. VRSTC: Occlusion-free video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7183–7192.

[22] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7132–7141.

[23] Shuping Hu, Kan Wang, Jun Cheng, Huan Tan, and Jianxin Pang. 2022. Triplet ratio loss for robust person re-identification. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Springer, 42–54.

[24] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. 2019. Global-local temporal representations for video person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 3958–3967.

[25] Jianing Li, Shiliang Zhang, and Tiejun Huang. 2019. Multi-scale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 8618–8625.

[26] Jianing Li, Shiliang Zhang, and Tiejun Huang. 2020. Multi-scale temporal cues learning for video person re-identification. IEEE Transactions on Image Processing 29 (2020), 4461–4473. DOI: 10.1109/TIP.2020.2972108

[27] Mengliu Li, Han Xu, Jinjun Wang, Wengpeng Li, and Yongli Sun. 2020. Temporal aggregation with clip-level attention for video-based person re-identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

[28] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 369–378.

[29] Zhaolu Li, Zongwei Zhou, Nan Jiang, Zhenjun Han, Junliang Xing, and Jianbin Jiao. 2020. Spatial preserved graph convolution networks for person re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications 16, 1s (2020), 1–14.

[30] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. 2018. Video-based person re-identification with accumulative motion context. IEEE Transactions on Circuits and Systems for Video Technology 28, 10 (2018), 2788–2802.

[31] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. 2019. Dense 3D-convolutional neural network for person re-identification in videos. ACM Transactions on Multimedia Computing, Communications, and Applications 15, 1s (2019), 1–19.

[32] Liangchen Liu, Xi Yang, Nannan Wang, and Xinbo Gao. 2020. Viewing from frequency domain: A DCT-based information enhancement network for video person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia. 227–235.

[33] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. 2021. Watching you: Global-guided reciprocal learning for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13334–13343.

[34] Yu Liu, Junjie Yan, and Wanli Ouyang. 2017. Quality aware network for set to set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5790–5799.

[35] Yiheang Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 8786–8793.

[36] Neeraj Matiyali and Gaurav Sharma. 2020. Video person re-identification using learned clip similarity aggregation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2655–2664.

[37] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. 2016. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1325–1334.

[38] Bo Pang, Deming Zhai, Junjun Jiang, and Xianming Liu. 2022. Fully unsupervised person re-identification via selective contrastive learning. ACM Transactions on Multimedia Computing, Communications, and Applications 18, 2 (2022), 1–15.

[39] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. 2021. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 783–792.

[40] Dripta S. Raychaudhuri and Amit K. Roy-Chowdhury. 2020. Exploiting temporal coherence for self-supervised one-shot video re-identification. In Proceedings of the European Conference on Computer Vision. Springer, 258–274.

[41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision. Springer, 17–35.
[42] Weijian Ruan, Chao Liang, Yi Yu, Zheng Wang, Wu Liu, Jun Chen, and Jiayi Ma. 2020. Correlation discrepancy insight network for video re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 4 (2020), 1–21.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 815–823.

[44] Chen Shen, Zhongming Jin, Wening Chu, Rongxin Jiang, Yaowu Chen, Guo-Jun Qi, and Xian-Sheng Hua. 2019. Multi-level similarity perception network for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2 (2019), 1–19.

[45] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. 2018. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[46] Arulkumar Subramaniam, Athira Nambiar, and Amurag Mittal. 2019. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 562–572.

[47] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*. 480–496.

[48] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1139–1147.

[49] Zengming Tang and Jun Huang. 2022. Harmonious multi-branch network for person re-identification with harder triplet loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 4 (2022), 1–21.

[50] Haoran Wang, Licheng Jiao, Fang Liu, Lingling Li, Xu Liu, Deyi Ji, and Weihao Gan. 2021. IPGN: Interactivity proposal graph network for human-object interaction detection. *IEEE Transactions on Image Processing* 30 (2021), 6583–6593. DOI: 10.1109/TIP.2021.3096333

[51] Haoran Wang, Licheng Jiao, Shuyuan Yang, Lingling Li, and Zexin Wang. 2020. Simple and effective: Spatial rescaling for person reidentification. *IEEE Transactions on Neural Networks and Learning Systems* 33, 1 (2020), 145–156. DOI: 10.1109/TNNLS.2020.3027589

[52] Hanzheng Wang, Jiaqi Zhao, Yong Zhou, Rui Yao, Ying Chen, and Silin Chen. 2021. AMC-net: Attentive modality-consistent network for visible-infrared person re-identification. *Neurocomputing* 463 (2021), 226–236. DOI: 10.1016/j.neucom.2021.08.053

[53] Kan Wang, Changxing Ding, Stephen J. Maybank, and Dacheng Tao. 2020. CDPM: Convolutional deformable part models for semantically aligned person re-identification. *IEEE Transactions on Image Processing* 29 (2020), 3416–3428. DOI: 10.1109/TIP.2019.2959923

[54] Kan Wang, Shuping Hu, Jun Cheng, Jiaxin Pang, and Huan Tan. 2022. RA loss: Relation-aware loss for robust person re-identification. In *Proceedings of the Asian Conference on Computer Vision*. 177–194.

[55] Kan Wang, Pengfei Wang, Changxing Ding, and Dacheng Tao. 2021. Batch coherence-driven network for part-aware person re-identification. *IEEE Transactions on Image Processing* 30 (2021), 3405–3418. DOI: 10.1109/TIP.2021.3060909

[56] Pengfei Wang, Changxing Ding, Zhiyin Shao, Zhbin Hong, Shengli Zhang, and Dacheng Tao. 2022. Quality-aware part models for occluded person re-identification. *IEEE Transactions on Multimedia* (2022). DOI: 10.1109/TMM.2022.3156282

[57] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2014. Person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2899–2908.

[58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2017. Non-local neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 7794–7803.

[59] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Exploit the unknown gradually: One-shot video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5177–5186.

[60] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 4733–4742.

[61] Sheng Xu, Chang Liu, Baochang Zhang, Jinhu Lü, Guodong Guo, and David Doermann. 2022. BiRe-ID: Binary neural network for efficient person re-ID. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 1s (2022), 1–22.

[62] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. 2020. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2899–2908.


[64] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 1–23.

[65] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. 2021. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 516–525.

[66] Wenyu Zhang, Qing Ding, Jian Hu, Yi Ma, and Mingzhe Lu. 2021. Pixel-wise graph attention networks for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5231–5238.

[67] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10407–10416.

[68] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. 2019. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4913–4922.

[69] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision*. Springer, 868–884.

[70] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J. Radke. 2019. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[71] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 1 (2017), 1–20.

[72] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13001–13008.

[73] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4747–4756.