sensors

MDPI

*Article*

# Multi-Granularity Aggregation with Spatiotemporal Consistency for Video-Based Person Re-Identification

**Hean Sung Lee [1], Minjung Kim [1], Sungjun Jang [1], Han Byeol Bae [2] and Sangyoun Lee [1,*]**

[1] School of Electrical and Electronic Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea; hslee2860@yonsei.ac.kr (H.S.L.); mjkima@yonsei.ac.kr (M.K.); jeu2250@yonsei.ac.kr (S.J.)

[2] School of Computer Science and Engineering, Kunsan National University, 558 Daehak-ro, Gunsan-si 54150, Republic of Korea; hbbae@kunsan.ac.kr

* Correspondence: syleee@yonsei.ac.kr; Tel.: +82-2-2123-5768

**Abstract:** Video-based person re-identification (ReID) aims to exploit relevant features from spatial and temporal knowledge. Widely used methods include the part- and attention-based approaches for suppressing irrelevant spatial–temporal features. However, it is still challenging to overcome inconsistencies across video frames due to occlusion and imperfect detection. These mismatches make temporal processing ineffective and create an imbalance of crucial spatial information. To address these problems, we propose the Spatiotemporal Multi-Granularity Aggregation (ST-MGA) method, which is specifically designed to accumulate relevant features with spatiotemporally consistent cues. The proposed framework consists of three main stages: extraction, which extracts spatiotemporally consistent partial information; augmentation, which augments the partial information with different granularity levels; and aggregation, which effectively aggregates the augmented spatiotemporal information. We first introduce the consistent part-attention (CPA) module, which extracts spatiotemporally consistent and well-aligned attentive parts. Sub-parts derived from CPA provide temporally consistent semantic information, solving misalignment problems in videos due to occlusion or inaccurate detection, and maximize the efficiency of aggregation through uniform partial information. To enhance the diversity of spatial and temporal cues, we introduce the Multi-Attention Part Augmentation (MA-PA) block, which incorporates fine parts at various granular levels, and the Long-/Short-term Temporal Augmentation (LS-TA) block, designed to capture both long- and short-term temporal relations. Using densely separated part cues, ST-MGA fully exploits and aggregates the spatiotemporal multi-granular patterns by comparing relations between parts and scales. In the experiments, the proposed ST-MGA renders state-of-the-art performance on several video-based ReID benchmarks (i.e., MARS, DukeMTMC-VideoReID, and LS-VID).

**Keywords:** video-based person re-identification; spatiotemporal learning; attention mechanism; complementary learning



**Citation:** Lee, H.S.; Kim, M.; Jang, S.; Bae, H.B.; Lee, S. Multi-Granularity Aggregation with Spatiotemporal Consistency for Video-Based Person Re-Identification. *Sensors* **2024**, *24*, 2229. https://doi.org/10.3390/s24072229

Academic Editor: Cosimo Distante

Received: 21 February 2024
Revised: 25 March 2024
Accepted: 27 March 2024
Published: 30 March 2024



## 1. Introduction

Person re-identification (ReID) is an essential application in large-scale surveillance systems and smart cities, aiming to identify individuals across different times and locations amidst varying conditions (e.g., camera views, occlusion, background clutter, illumination, scale, and body pose). With the growth of video surveillance systems, advanced video ReID methods [1–8] have been attracting attention due to their potential to offer larger capacity for achieving more robust performance. In contrast to image-based ReID, which relies solely on a single image, video-based approaches harness a richer source of temporal information. Consequently, most video-based methods [6–14] predominantly concentrate on feature extraction and aggregating such spatiotemporal knowledge.

Previous methods commonly fall into two categories: part-based [14–17] and attention-based approaches [4,7,11,13,18–20]. These methods segment global features into partial

information and then derive a single feature vector by leveraging relationships among relevant spatial and temporal knowledge. Specifically, the methods [14,15] spatially separate images or global features into fixed partitions. Ref. [14] employed horizontal separation and utilized the graph convolutional network (GCN) [21] to aggregate spatial and temporal dimensions. Similarly, Ref. [15] employed horizontal separation at multiple scales to divide details and aggregate them through hypergraphs at various granularity levels. Despite these efforts, challenges persist due to temporal misalignment caused by object occlusion or inaccurate detection algorithms during feature aggregation. When features are separated into horizontal parts, they inconsistently include unnecessary information along the temporal axis during occlusion or detection errors (Figure 1a). Such inconsistencies, particularly in video ReID, potentially lead to interference in features and result in inaccurate outcomes. Alternatively, attention mechanisms such as [22–25] have been widely utilized to enhance feature representation by accentuating relevant regions while suppressing irrelevant areas. Recent video-based ReID approaches [18,19] have explored attention-based partitioning methods to leverage diverse attention parts. However, these methods typically create sub-attentions separate from the fixed main attention, resulting in imbalanced information across parts and restricting the number of semantic part divisions. They may tend to prioritize parts with abundant information, potentially overlooking finer details of targets with relatively lesser information. This could result in inaccurate outcomes when crucial parts of the target are occluded. To overcome the above problem, we aim to extract enhanced ReID features by fully exploiting detailed information in the spatiotemporal information by ensuring uniform information quantity across parts and maintaining consistent semantic information temporally.

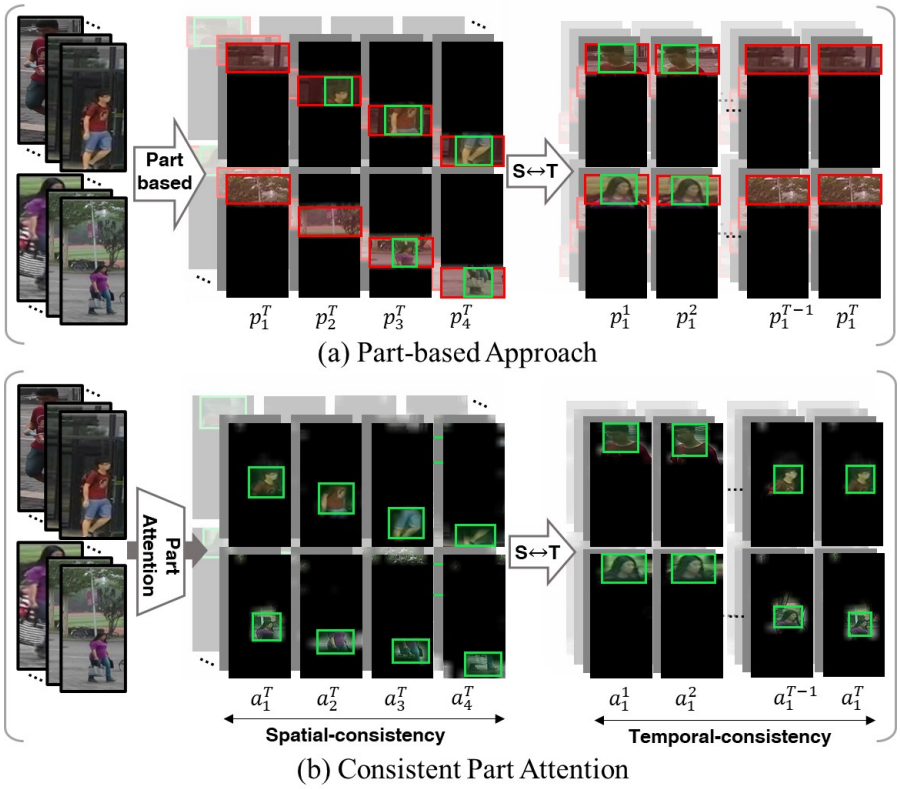


**Figure 1.** Comparison of the (**a**) part-based approach and (**b**) proposed consistent part-attention (CPA) method. As shown in the figure, the part-based approach has a different amount of relevant spatial information with some interferences for each part, whereas the CPA method provides only key information uniformly. When temporal misalignment occurs, the part-based approach demonstrates inconsistent human parts. In contrast, CPA consistently offers semantic information about the same part.

In this paper, we introduce the *consistent part-attention* (CPA) module, which effectively manages uniform spatiotemporal information without interference or noise. Notably, CPA learns uniform attention in spatiotemporal dimensions solely through self-information and a few priors, eliminating the need for hard labels such as human parsing or skeleton data. As illustrated in Figure 1b, the CPA module not only eliminates interference and noise in spacetime, but also ensures consistent delivery of semantic information to the model, averting uneven information distribution and ensuring the thorough capture of fine target details.

Addressing the challenge of video ReID entails leveraging both spatial and temporal knowledge across various granularities. To this end, we employed the *Multi-Attentive Part Augmentation* (MA-PA) scheme to obtain multi-granularity information with various attention scales. Multi-granularity information [15,26] has shown promise in incorporating detailed features in videos. The MA-PA generates multi-granularity attention by recombining fine attention from CPA. As shown in Figure 2a, merging segmented part information can alter and diversify semantic meanings (i.e., when the target is small, distinguishing facial or footwear details at a smaller scale becomes challenging; however, combining these with parts related to shirts or pants extends semantic information to upper and lower body regions). This empowers the model to capture robust information across a spectrum of semantic meanings, from fine-grained to broader details.

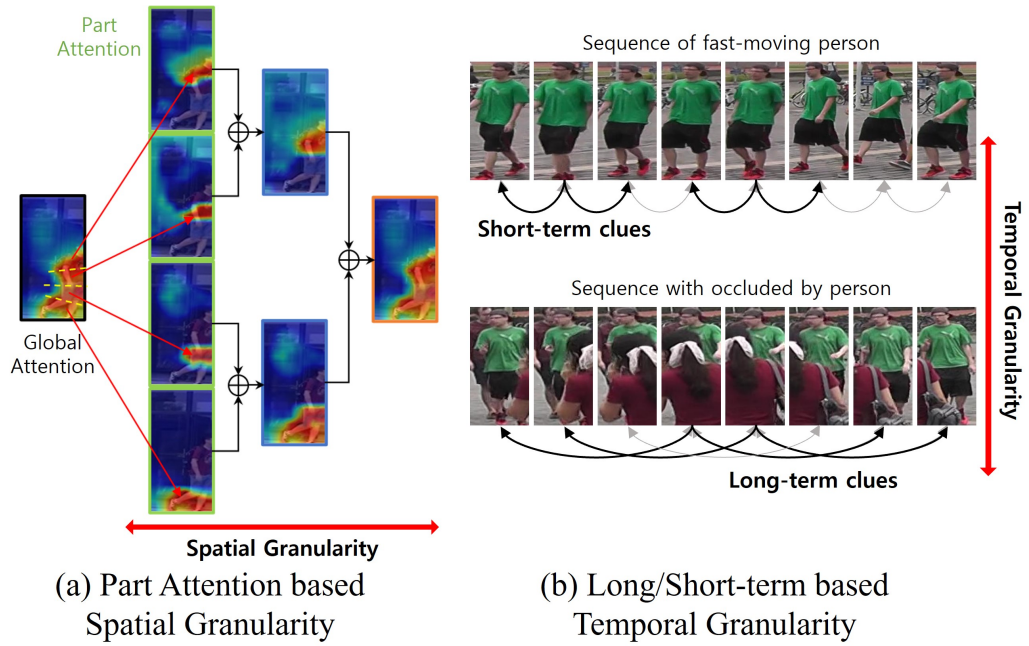


**Figure 2.** Illustration of spatiotemporal multi-granularity. (**a**) Part-attention-based spatial granularity synthesizes diverse semantic attention at multiple granular levels by combining sub-part attentions. (**b**) Long-/short-term-based temporal granularity comprises granularities with varying sampling intervals to adopt the context of different sequences for the long and short terms.

To capture temporal relations, we employed the *Long-/Short-term Temporal-Augmentation* (LS-TA) module, which obtains multi-granularity temporal information. LS-TA conducts time sampling at different intervals to harness the overall temporal advantage. Long- and short-term temporal cues have been utilized for temporal modeling due to their respective crucial patterns [19,27,28]. As shown in Figure 2b, varying sampling intervals yield distinct features. For instance, short-term clues reflect the target's motion patterns, while long-term cues effectively alleviate occlusion. Consequently, LS-TA yields diverse temporal features, enabling the model to extract robust features in various situations. After augmenting spatial and temporal granular cues, we propose the *Spatiotemporal Multi-Granularity Aggregation* (ST-MGA) to exploit densely separated spatial and temporal clues simultaneously. ST-MGA

investigates the relations between multi-granular and part information from both spatial and temporal cues. Since the granular part features refined in the previous process contain all the information in spatiotemporal dimensions without interference, ST-MGA can extract robust and complementary features in any situation.

To summarize, our main contributions are as follows. We designed a *consistent part-attention* (CPA) module to provide spatiotemporally consistent and well-aligned part attention. To exploit the multi-scale granularity, we introduce *Multi-Attention Part Augmentation* (MA-PA), which uses the fine part features from CPA to synthesize semantic parts at multiple scales spatially. We also suggest *Long-/Short-term Temporal Augmentation* (LS-TA) for considering relations at various temporal scales. The temporally consistent part information through CPA allows LS-TA to have the full advantage of temporal knowledge. Using the spatiotemporal multi-granularity part information, we propose *Spatiotemporal Multi-Granularity Aggregation* (ST-MGA), which performs partwise and scalewise aggregation. The ST-MGA method investigates the relations between multi-granular and part information from both spatial and temporal cues and encourages complementary features for video person ReID. In the experiments, we validated the effectiveness of our approach on multiple benchmarks. Our approach outperforms previous state-of-the-art methods on several video ReID benchmarks and shows more accurate attention parts than the existing part-based approaches.

## 2. Related Work

### 2.1. Video-Based Person ReID

In recent years, video-based person ReID [7,14,15,17–20,26–43] has garnered significant attention due to the abundant temporal and spatial cues available in videos. The predominant approach in video ReID is extracting and aggregating dynamic spatiotemporal features. Some methods employ recurrent architectures [5,6,44] for video representation learning to leverage temporal cues. Refs. [28,45] utilized 3D-convolution [46,47] for spatial–temporal feature learning. A temporal attention mechanism [8,9,20,48] has also been proposed for robust temporal feature aggregation. In recent research, to contain richer temporal and spatial information, many methods [20,39–42] have been proposed. Ref. [39] presented a statistic attention (SA) block to capture long-range high-order dependencies of the feature maps. Ref. [40] used hierarchical mining, which mines the characteristics of pedestrians by referring to the temporal and intra-class knowledge. Ref. [41] proposed a saliency and granularity mining network to learn the temporally invariant features. Ref. [42] implemented a two-branch architecture to separately learn the pose feature and appearance feature and concatenated them together for more discriminative representation. Ref. [20] removed interference and obtained key pixels and frames by learning attention-guided interference-removal modules. A simple literature survey of the previous methods is shown in Table 1.

Recently, refs. [7,26] focused on aggregating diverse partial information, both spatially and temporally. To obtain partial spatial cues, certain approaches have adopted horizontal partitioning [14–17] or explored diverse attention mechanism [7,11,18,19]. However, most of these methods cannot fully exploit the potential of spatiotemporal knowledge. Horizontal partitioning often struggles to maintain information consistency in cases of temporal misalignment due to an inaccurate detector. The diverse attention mechanisms have an unbalanced information distribution regarding attention, leading to inefficient aggregation. To exploit the full advantage of spatiotemporal information, we first propose a straightforward, yet effective framework, called *consistent part attention* (CPA), designed to ensure the consistency of partial information and lead to efficient aggregation in the spatial and temporal dimensions. Then, we efficiently the aggregate spatiotemporal partial information using the *Spatiotemporal Multi-Granularity Aggregation* (ST-MGA) scheme to extract complementary video features.

**Table 1.** Methodology of video-based person re-identification.

| Reference | Source | Methodology |
|---|---|---|
| M3D [28] | AAAI'19 | Proposed multi-scale 3D-convolution layer to refine the temporal features |
| STA [7] | AAAI'19 | Proposed spatial–temporal attention approach to fully exploit discriminative parts |
| GLTR [27] | ICCV'19 | Proposed global–local temporal representation to exploit multi-scale temporal cues |
| RGSA [29] | AAAI'20 | Designed relation-guided spatial-attention module to explore discriminative regions |
| FGRA [30] | AAAI'20 | Proposed frame-guided region-aligned model to extract well-aligned part features |
| MG-RAFA [26] | CVPR'20 | Suggested attentive feature aggregation with multi-granularity information |
| PhD [31] | CVPR'20 | Proposed Pompeiu–Hausdorff distance learning to alleviate the data noise problem |
| STGCN [14] | CVPR'20 | Jointly optimized two GCN branches in spatial and temporal dimensions for complementary information |
| MGH [15] | CVPR'20 | Designed a multi-granular hypergraph structure to increase representational capacities |
| TCLNet [18] | ECCV'20 | Introduced a temporal-saliency-erasing module to focus on diverse part information |
| AP3D [32] | ECCV'20 | Proposed appearance-preserving 3D-convolution to align the adjacent features at the pixel level |
| AFA [33] | ECCV'20 | Proposed adversarial feature augmentation, which highlights the temporal coherence features |
| SSN3D [34] | AAAI'21 | Designed a self-separated network to seek out the same parts in different frames |
| BiCnet-TKS [19] | CVPR'21 | Used multiple parallel and diverse attention modules to discover diverse body parts |
| STMN [35] | ICCV'21 | Leveraged spatial and temporal memories to refine frame-/sequence-level representations |
| STRF [36] | ICCV'21 | Proposed spatiotemporal representation factorization for learning discriminative features |
| SINet [37] | CVPR'22 | Designed SINet to enlarge attention regions for consecutive frames gradually |
| CAVIT [38] | ECCV'22 | Used a contextual alignment vision transformer for spatiotemporal interaction |
| SANet [39] | TCSVT'22 | Introduced the SA block, which can capture long-range and high-order information |
| HMN [40] | TCSVT'22 | Designed hierarchical mining network which can mine as many characteristics |
| SGMN [41] | TCSVT'22 | Designed a saliency- and granularity-mining network for discovering temporal coherence |
| BIC+LGCN [42] | TCSVT'23 | Used a branch architecture to separately learn appearance features and human pose |
| IRNet-V [20] | TCSVT'23 | Proposed an interference-removal framework for removing various interferences |

### *2.2. Attention for Person ReID*

The attention mechanism, as discussed in [22–25], has been widely used in person ReID to enhance representation by emphasizing the relevant features and suppressing irrelevant ones. In image-based ReID, Refs. [49–58] learned attention in terms of the spatial or channel dimension. In some studies of video-based ReID, temporal attention is performed to weigh and aggregate frame-level features [5,9,59]. Moreover, Ref. [26] proposed joint spatial and temporal attention to exploit relations at multiple granularities. Recently, Refs. [18,19] proposed diverse spatial-attention modules to enhance video representation. The diverse attention modules focus on different regions for consecutive frames. However, they create sub-attention parts separate from the attention of the main frame, which restricts the number of semantic attention parts and results in each part containing an inconsistent amount of information. This limitation leads to inefficiencies in aggregating diverse features because the focus remains on the main attention part. Unlike the above methods, the proposed CPA provides uniform spatial information and temporally coincident cues about the diverse attention parts, leading to efficient aggregation in the spatial and temporal dimensions.

### *2.3. Spatiotemporal Aggregation*

Capturing spatial and temporal information is critical to learning comprehensive representations of videos effectively. The most-used approach [4,18,27,32,34] involves using convolutional neural networks (CNNs) to extract spatial features from individual video frames and integrating these features with temporal modeling. Various methods, such as RNNs, 3D-CNNs, GCNs, and attention mechanisms, can be employed for spatiotemporal aggregation. Ref. [28] introduced a compact 3D convolutional kernel that facilitates multi-scale temporal feature learning by incorporating long-term temporal modeling and refining appearance features through spatiotemporal masking. Ref. [19] focused on capturing visual features by considering the spatial details and long-distance context information, which are combined using a multi-scale temporal kernel in the 3D convolutional layers. Ref. [60] proposed graph convolution to directly propagate cross-spacetime and cross-scale information, capturing high-order spatial–temporal correlations. Ref. [35] addressed the problem of spatial distractors by memorizing them and suppressing distracting scene details while using temporal attention patterns to aggregate the frame-level representation.

Ref. [61] learned spatiotemporal information attention using ConvLSTM [62] to explicitly capture and aggregate spatiotemporal information in video-based industrial smoke emission recognition. Ref. [63] explored spatial correlations within each frame to determine the attention weight of different locations and also considered temporal correlations between adjacent frames.

## 3. Methodologies

In this section, we propose *Spatiotemporal Multi-Granularity Aggregation* (ST-MGA) methods with the proposed spatiotemporally consistent cues. We introduce the preliminaries in Section 3.1. Then, we describe the proposed *consistent part-attention* (CPA) module, which aims to provide spatiotemporally consistent information by using a simple, but effective attention approach, in Section 3.2. For multi-granular spatial and temporal information, we introduce *Multi-Attention Part Augmentation* (MA-PA) and *Long-/Short-term Temporal Augmentation* in Section 3.3. Last, we present ST-MGA for spatiotemporally complementary features in Section 3.4.

### *3.1. Overview*

To improve video-based person ReID, we aimed to extract consistent spatial and temporal cues, and aggregated this information for complementary feature extraction. The overall framework of the proposed approach is illustrated in Figure 3. As the input video clip, we randomly sampled $T$ frames as $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$. Then, we extracted video features $F \in \mathbb{R}^{T \times C \times H \times W}$ through the CNN backbone (e.g., ResNet-50 [64] pretrained on ImageNet [65]), where $H$, $W$, and $C$ represent the height, width, and number of channels, respectively. To extract the part cues, we employed the CPA module, which ensures that extracted attentions contain consistent spatiotemporal semantic information. Using CPA, we obtained $N_s$ sub-attentions, $A = \{a^n | n = 1, 2, \cdots, N_s\}$. Subsequently, we extracted multi-granular parts using the MA-PA method. MA-PA augments multi-granular attention with varying scales by utilizing attention parts from CPA. Augmented multi-granular attention was applied to $F$ to generate $N_P$ granular features. To consider temporal relations, we also jointly augmented $N_t$ long- and short-term features using LS-TA by combining multi-scale temporal cues. We define part features after MA-PA and LS-TA as $\hat{P}$ and $\tilde{P}$, respectively. Leveraging $N_p \times N_t$ multi-granular features, we propose ST-MGA, which exploits spatial and temporal relations and aggregates different levels of part features. Finally, after partial and temporal averaging of the aggregated features, we extracted the complementary video features, $Z_c \in \mathbb{R}^C$. The notations and their corresponding descriptions are presented in Table 2.

**Table 2.** Notations and descriptions.

| Notations | Descriptions |
|---|---|
| $\mathcal{V}$ | Input video |
| $T$ | Length of $\mathcal{V}$ |
| $F$ | Output feature from backbone |
| $A$ | Set of part attentions from CPA |
| $\dot{A}$ | Set of residual part attentions in CPA |
| $P$ | Set of part-attentive features |
| $\hat{A}$ | Set of augmented part attentions from MA-PA |
| $\hat{P}$ | Set of augmented spatial attentive features after MA-PA |
| $a'$ | Temporal attention value in LS-TA |
| $\tilde{P}$ | Set of augmented multi-granular features after LS-TA |
| $Z^F$ | Global averaged feature from backbone |
| $Z^T$ | Temporal aggregated feature from T-MGA |
| $Z^C$ | Final complementary video feature from S-MGA |

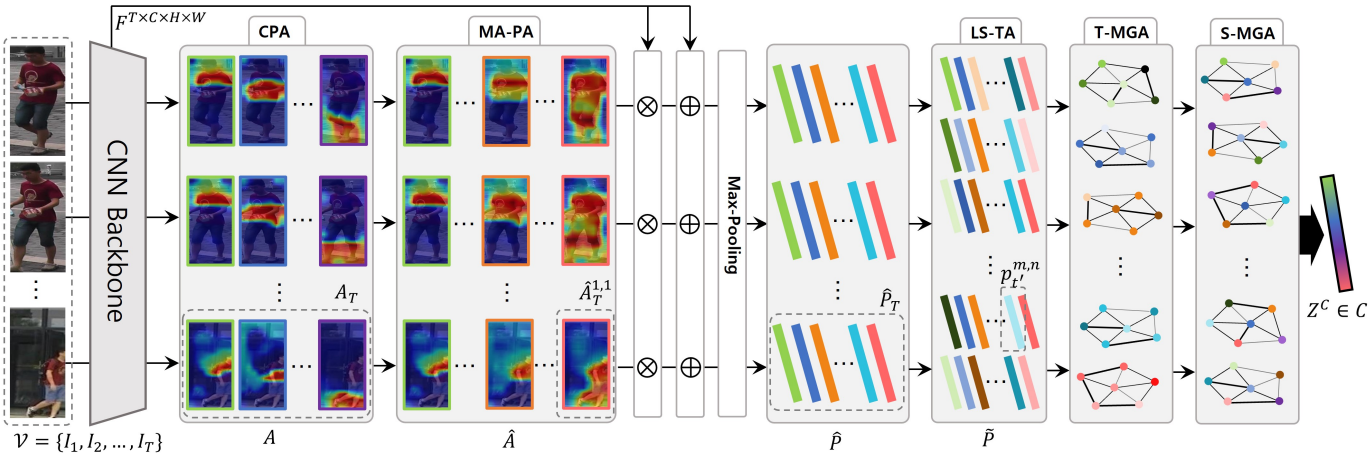


**Figure 3.** Overview of our framework. We sampled $T$ frames in the video sequence and extracted features $F$ from the CNN backbone. We first extracted sub-attentions $A$ through the CPA module with spatiotemporally consistent information. For multiple spatial granularities, we augmented part attention $\hat{A}$ by combing $A$ hierarchically. Subsequently, we extracted the part-attentive features $\hat{P}$ by applying $\hat{A}$ and the max-pooling operation to $F$. We jointly augmented $\hat{P}$ to $\tilde{P}$, which contain long-/short-term features, to consider temporal relations. Leveraging the augmented-spatiotemporal-part features, we exploited all video cues by aggregating the part features spatially (S-MGA) and temporally (T-MGA). Last, we extracted the complementary video person ReID features.

### *3.2. Spatiotemporally Consistent Part Attention*

Given the $t$-th global features $F_t \in \mathbb{R}^{H\times W\times C}$, we first optimized the global attention through a learnable model. We extracted the global attention value $a_t^g$ for the $t$-th frame from the global-attention module $GA$ composed as follows:

$$a_t^g = Sigmoid(W_2 ReLU(W_1 F_t)), \tag{1}$$

where $W_1 \in \mathbb{R}^{\frac{C}{\gamma}\times C}$ and $W_2 \in \mathbb{R}^{1\times\frac{C}{\gamma}}$ are implemented by $1\times 1$ convolution with shrink ratio $\gamma$ followed by BN. Subsequently, we can extract global attentive features $P_g \in \mathbb{R}^C$ as:

$$P_g = \frac{1}{T}\sum_{t=1}^{T}(\mathcal{P}(F_t + a_t^g \odot F_t)), \tag{2}$$

where the symbol $\odot$ represents elementwise multiplication and $\mathcal{P}$ is the global max-pooling operation. To optimize $a^g$, we applied the batch hard triplet loss [66] and softmax cross-entropy loss with $P_g$ as the input. The two loss formulas are denoted as $\mathcal{L}_{tri}^g$ and $\mathcal{L}_{CE}^g$.

After obtaining $a^g$, we optimized the CPA by using $a^g$. As illustrated in Figure 4, CPA composes $N_s$ part-attention modules $PA$, which extract $N_s$ different semantic sub-attentions $A_t = \{a_t^n\}_{n=1}^{N_s}$. Each $PA$ module has the same structure as $GA$, with different parameters. To encourage each attention focus to have different semantic information, we define the following priors: (i) the features applied by each attention must be semantically classifiable for each other; (ii) the sum of all sub-attention parts should be global attention; (iii) each sub-attention part should contain a uniform amount of information.

Based on the first prior, we designed a part classifier that classified $N_s$ parts as the class ID. We first extracted the part-attention features $P = \{P_n\}_{n=1}^{N_s}$ in the same way as $P_g$ with the corresponding attention. Then, we trained each attentive feature to be classifiable by setting a different label $y^n$, a one-hot vector for the $n$-th attentive part. To this end,

we employed the softmax cross-entropy loss with input $P_n$. We define this loss formula as follows:

$$\mathcal{L}_{PCE} = \mathbb{E}[-log(p(y^n|P_n))], \tag{3}$$

where $p(y^n|P_n)$ is the predicted probability that input $P_n$ belongs to its part label $y^n$. $\mathcal{L}_{PCE}$ encourages each attention to focus on different relevant areas because each attentive feature must be semantically classifiable despite being applied to the same input features.

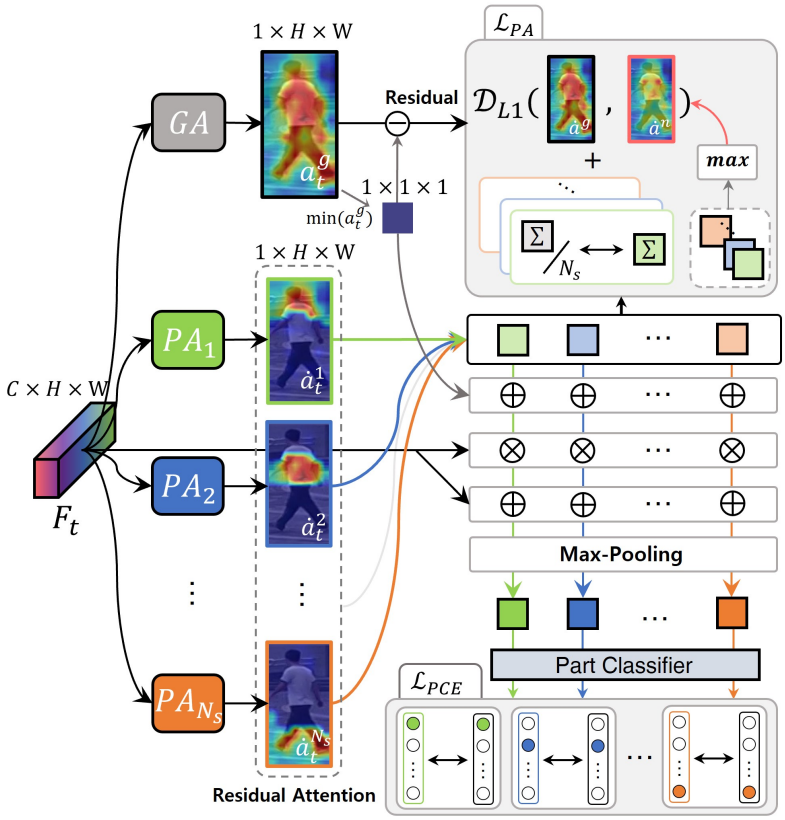


**Figure 4.** Learning framework for the consistent part-attention (CPA) module. The CPA includes one global-attention module, $GA$, to extract global attention $a^g$, $N_s$ part-attention modules, $PA$, and one part classifier, $\mathcal{C}_P$. In CPA, we used $a^g$ to encourage each attention part to be optimized based on predefined priors.

Before applying the second and third priors, we introduced a technique to extract precise part-attention information. The technique is that sub-attentions only learn the residual from the background value of the global attention when directly inducing each attention part. The rationale behind this approach is to focus on the relevant region only, rather than the background. If background information is not excluded, arbitrary attention is slightly highlighted on the background region as the interest, making interference during aggregation. The goal is for each attention part to have equal importance; thus, we proceeded with learning by extracting only the residuals from the background values of the global attention. By excluding the background information, attention is directed to the relevant areas only. We analyze this in Section 4.

As shown in Figure 4, we first define the minimum of $a^g$ of the background score and obtained the residual global attention, $\dot{a}^g$, by removing it. We define the residual part attentions as $\dot{A} = \{\dot{a}^n\}_{n=1}^{N_s}$, which are learned by focusing only on the semantic area and ignoring unnecessary parts (e.g., the background and occlusions). We utilized $\dot{a}$ and $\dot{a}^g$ to impose constraints on the second and third priors. First, we designed the $L1$ distance loss, which sets the sum of all residual part attentions $\dot{A}$ equal to $\dot{a}^g$. Additionally, we

encouraged the spatial sum of each residual part attention to match the spatial sum of $\tilde{a}^g$ divided by $N_s$. We combined the first and second loss formulas as follows:

$$\mathcal{L}_{PA} = || \sum_{n=1}^{N_s} \dot{a}^n - \dot{a}^g || + \sum_{n=1}^{N_s} || \sum_{k=1}^{K} \dot{a}_k^n - \frac{1}{N_s} \sum_{k=1}^{K} \dot{a}_k^g ||, \tag{4}$$

where $K = H \times W$. $\mathcal{L}_{PA}$ simply, but strongly satisfies the prior we predefined. Furthermore, $\mathcal{L}_{PA}$ encourages an emphasis on the position of relevant and uniform importance in each part.

The overall loss function for training CPA is formulated as follows:

$$\mathcal{L}_{CPA} = \mathcal{L}_{tri}^g + \mathcal{L}_{CE}^g + \mathcal{L}_{PA} + \lambda \mathcal{L}_{PCE}, \tag{5}$$

where $\lambda$ is a scaling parameter for weighting $\mathcal{L}_{PCE}$. The attention parts extracted using CPA represent temporally consistent semantic cues with uniform spatial importance. We leveraged these attention parts instead of the previous simple horizontal partitions to take full advantage of this in the video. In subsequent sessions, we will denote $\tilde{A}$ as $A$ for convenience.

### *3.3. Multi-Granularity Feature Augmentation*

To further improve video feature representation, we augmented diverse spatial and temporal cues at multi-granular levels. With the efficient CPA module to obtain consistent part attention $A$, we applied Multi-Attentive Part Augmentation (MA-PA), which uses $A$ as a sub-attention to generate the parent attention. Following [15], we hierarchically combined the sub-attention into a granular scale $m \in \{0, 1, 2, \cdots, M\}$. To avoid excessive duplication of certain parts, we prevented overlapping when combining sub-attentions, as shown in Figure 2a. The parent granular attention comprises the non-duplicated combination of child sub-attentions. We define the $t$-th augmented part attention as $\hat{A}_t = \{A_t^m | m = 0, 1, 2, \cdots, M\}$, where $\hat{A}_t^m = \{\hat{a}_t^{m,n} | n = 1, 2, \cdots, 2^m\}$. $\hat{a}_t^{m,n}$ represents attention to the $n$-th part of the $m$-th scale, and each scale has $2^m$ part attentions. Subsequently, we obtained $N_P = \sum_{m=0}^{M} 2^m$ attentive features $\hat{P}$ in each frame as follows:

$$\hat{p}_t^{m,n} = \mathcal{P}(F_t + \hat{a}_t^{m,n} \odot F_t), \tag{6}$$

where $\hat{p}_t^{m,n} / \hat{a}_t^{m,n}$ is the $n$-th part feature/attention with scale $m$ and $\mathcal{P}$ is the global max-pooling operation. As with $\hat{P}$, each part feature at the same granular level contains uniform semantic information.

Using augmented spatial attentive features $\hat{P}$, we employed Long-/Short-term Temporal Augmentation (LS-TA) to augment long- and short-term features jointly. As shown in Figure 2b, the importance of long- and short-term temporal relations can vary depending on the sequence context. However, simple temporal processing is often inefficient due to temporal inconsistencies such as misalignment between adjacent frames [32]. In our study, we have already addressed these temporal inconsistency problems with the proposed CPA; therefore, we can proceed with the temporal granularity aggregation without concern. To this end, we first applied temporal attention from each frame level by predicting the framewise scores as follows:

$$\tilde{p}_t = \hat{p}_t + a'_t \odot \hat{p}_t, \tag{7}$$

$$a'_t = Sigmoid(W'_2 ReLU(W'_1 hayp_t)), \tag{8}$$

where $a'_t$ is the temporal attention value, $W'_1 \in \mathbb{R}^{\frac{C}{\gamma} \times C}$ and $W'_2 \in \mathbb{R}^{1 \times \frac{C}{\gamma}}$ are implemented by 1D-convolution with a kernel size of 1, and $\gamma$ is the shrink ratio. Then, following [19], we conducted a temporal-select operation $\{\mathcal{S}^{t'} : \hat{P} \rightarrow \tilde{P}_{t'} \in \mathbb{R}^{T \times C}\}$, where $\mathcal{S}$ is the frame-select operation of a 1D-temporal convolution [67] with a kernel size of $2t' + 1$ and $t' \in 1, 2, \cdots, T'$

is the temporal granular scale. We varied $t'$ to obtain multiple temporal granular features at different intervals from short to long term. With LS-TA, we generated $N_T = T \times T'$ temporal granular features in each part and combined them with the granular part, and a total of $N_P \times N_T$ spatiotemporal granular cues were stored. We leveraged these granular cues to exploit the spatiotemporal coverage in the video.

### *3.4. Spatiotemporal Multi-Granularity Aggregation*

To extract optimal video person ReID features, it is important to effectively aggregate the multi-granular information from the spatial and temporal information. From this view, we propose a Multi-Granularity-Aggregation (MGA) module that fully exploits the spatiotemporal cues. The structure of MGA is illustrated in Figure 5. MGA primarily comprises partwise and scalewise aggregators. The partwise aggregator has $M$ parallel branches with the same structure, where $M$ is the simplified scale number. For each scale of the granularity level, we represent $N$ part features as $\tilde{p}_i^m \in \mathbb{R}^{N \times C}$, where $i = \{1, \cdots, N\}$ and $N$ is the simplified number of parts. The partwise relation matrix in the $m$-th scale $R^m \in \mathbb{R}^{N \times N}$ can be defined as a dot-product affinity as follows:

$$R_{i,j}^m = \frac{\exp(\varphi(\tilde{p}_i^m)^T \omega(\tilde{p}_j^m)/\tau)}{\sum_{k=1}^{N} \exp(\varphi(\tilde{p}_i^m)^T \omega(\tilde{p}_k^m)/\tau)}, \tag{9}$$

where $\varphi$ and $\omega$ are linear functions and $\tau$ is the temperature hyperparameter. With the relation matrix $R^m$, we calculated partwise aggregated features as $\phi(\tilde{p})^T R_m$, where $\phi$ is a linear function. Then, the output features were extracted using average pooling of all part cues, followed by the residual addition. After the partwise aggregator, we concatenated $M$ outputs, denoted as $Y \in \mathbb{R}^{M \times C}$.

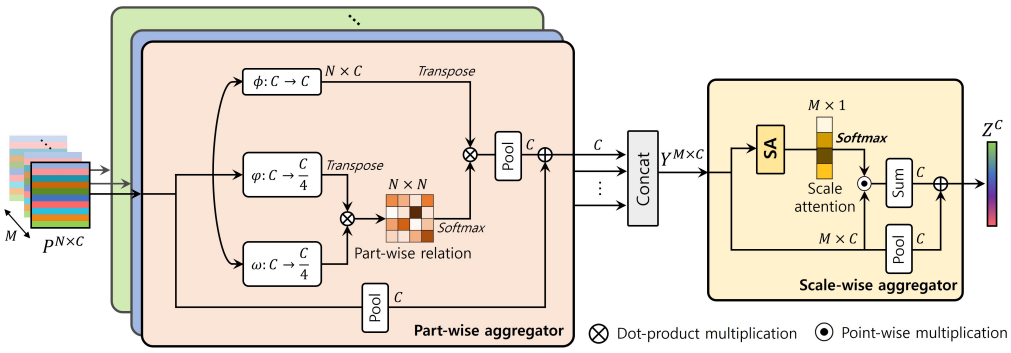


**Figure 5.** The architecture of Multi-Granularity Aggregation (MGA). As the input $\tilde{P}$, partwise aggregation proceeds first at each granular scale. Then, the granular features $X$ are aggregated using a scale-attention block. Last, MGA extracts complementary video ReID features, $Z$. Temporally, $N$ and $M$ are changed to $T$ and $T'$, respectively.

With input $Y$, scalewise aggregation proceeds using a scalewise aggregator. To aggregate all granular scale cues on $Y$, we designed a scale-attention (SA) module that predicts the scalewise attention score to weigh the aggregation. We normalized the learned attention scores from the SA module via the softmax function across scale dimensions and obtained the scale-attention score $A_s \in \mathbb{R}^{C \times M}$:

$$A_s = SA(Y) = Softmax(W_{\theta_2}(ReLU(W_{\theta_1} Y))/\tau), \tag{10}$$

where $W_{\theta_1} \in \mathbb{R}^{\frac{C}{\eta} \times C}$ and $W_{\theta_2} \in \mathbb{R}^{1 \times \frac{C}{\eta}}$ are a fully connected layer with shrink ratio $\eta$ and $\tau$ is the temperature hyperparameter. Afterward, we extracted complementary video representation $Z \in \mathbb{R}^C$ using scalewise attention $A_s$ to aggregate all scale features as follows:

$$Z = \sum_{m=1}^{M} A_s \odot Y + \mathcal{A}(Y). \tag{11}$$

We applied MGA separately in the spatial (S-MGA) and temporal (T-MGA) information. S-MGA aggregates partwise and scalewise based on spatially attentive-part features, whereas T-MGA is based on the temporal-part features. As illustrated in Figure 5, T-MGA has the same structure as S-MGA, with $N$ and $M$ replaced by $T$ and $T'$, respectively. By applying T-MGA and S-MGA, we extracted final video features $Z^C$, enhanced by exploiting the spatiotemporal cues across each scale.

For the reliable learning of each stage, we used three types of features $Z' = \{Z^F, Z^T, Z^C\}$, where $Z^F$ is the global features of the backbone, $Z^T$ represents the temporally aggregated features from T-MGA, and $Z^C$ denotes the final complementary video features from S-MGA. The overall joint objective is defined as follows:

$$\mathcal{L}_{total} = \sum_{z \in Z'} (\mathcal{L}_{CE}^{z} + \mathcal{L}_{tri}^{z}) + \mathcal{L}_{CPA}, \tag{12}$$

where $\mathcal{L}_{CE}$ is the softmax cross-entropy loss and $\mathcal{L}_{tri}$ is the batch-hard triplet loss [66]. In the inference phase, we only used the complementary video person ReID features, $Z_C$.

## 4. Experiments

In this section, we extensively evaluate the proposed framework in video-based person ReID scenarios. We conducted comprehensive experiments on three benchmark datasets, and provide a detailed analysis of the results. Additionally, we included extensive ablation studies to investigate the effectiveness of our approach further.

### *4.1. Datasets and Evaluation Metric*

We evaluated the proposed framework on three challenging video ReID datasets: MARS [2], DukeMTMC-VideoReID (Duke-V) [68], and LS-VID [27]. MARS is one of the large-scale benchmark datasets for video ReID, which contains 17,503 tracklets of 1261 identities and an additional distractor of 3248 tracklets. There are substantial bounding box misalignment problems to make it more challenging. Duke-V is a widely used large-scale video ReID dataset captured by 8 cameras with 4832 tracklets of 1404 identities. LS-VID is the most recent large-scale benchmark dataset for video ReID. It contains 3772 identities and 14,943 tracklets captured by 15 cameras. There are many challenging elements, such as varying illumination and bounding box misalignment, to make it close to a real-world environment, A summary comparison is illustrated in Table 3. For the evaluation, we used only the final complemented video ReID features $Z_C$ during the inference stage. Moreover, we assessed the performance using Rank-1, Rank-5, and Rank-10 accuracy for cumulative matching characteristics (CMCs) and the mean average precision (mAP). Rank-k and the mAP are the most popular evaluation metrics for person ReID. Rank-k measures the accuracy by evaluating whether the correct match appears within the top-k-ranked results. To this end, for each query, an algorithm will rank all the gallery samples according to their distances. The mAP evaluates how well the system ranks the retrieved matches for each query. It considers both precision and recall, providing a comprehensive assessment of the re-identification system's performance across different query scenarios.

**Table 3.** Statistics between different video ReID datasets.

| | **MARS** | **Duke-V** | **LS-VID** |
|---|---|---|---|
| # Identities | 1261 | 1404 | 3772 |
| # of Videos | 20,751 | 4832 | 14,943 |
| # of Cameras | 6 | 8 | 15 |
| B-Box | DPM | manual | FRCNN |

### *4.2. Implementation Details*

Following the common practices in ReID [19,27,32,69], we used ResNet-50 [64] trained by ImageNet classification [65] as the backbone of the proposed model for a fair and effectiveness validation. Similar to [69,70], we removed the last down-sampling operation to enrich granularity, resulting in a total down-sampling ratio of 16. We employed a random sampling strategy to select 8 frames for each video sequence with a stride of 4 frames. Each batch contains 16 identities, each with 4 tracklets. The input frames were resized to $256 \times 128$. Random horizontal flip and random erasing [71] were used for data augmentation. The training process utilizes the Adam [72] optimizer for 150 epochs. The learning rate was initialized to $3.5 \times 10^{-4}$ and decayed every 40 epochs using a decay factor of 0.1. For the hyperparameters of the proposed framework, a greedy search was conducted about $\lambda$ and $\tau$ to increase reproducibility and transparency. $\lambda$ is the weight for $\mathcal{L}_{PCE}$ of Equation (5), and $\tau$ is the temperature hyperparameter, which indicates sensitivity to the relation of Equations (9) and (10). As shown in Figure 6, we set $\lambda$ to 0.1 and $\tau$ to 0.05 as the optimal parameters. The shrink ratios $\gamma$ and $\eta$ were set to 16 and 4, respectively. The framework was implemented with *PyTorch1.4* [73] on one GeForce RTX 3090 (Nvidia, Santa Clara, CA, USA).

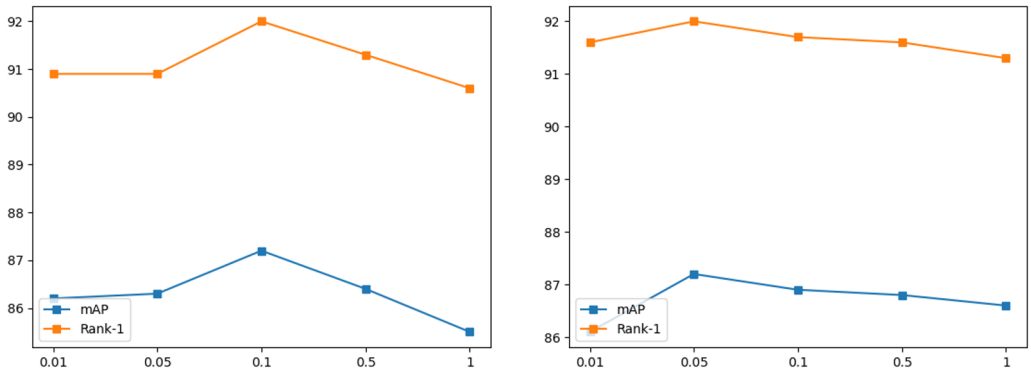


**Figure 6.** Sensitivity analysis on hyperparameters for (**a**) $\lambda$ and (**b**) $\tau$ in MARS [2].

### *4.3. Ablation Study*

#### 4.3.1. The Influence of CPA

We first evaluated the effect of the proposed CPA module with different components of loss functions and techniques. To extract spatiotemporally consistent information, we suggest three priors and approaches with $\mathcal{L}_{PCE}$, $\mathcal{L}_{PA}$, and a residual technique (*res*). We conducted only spatial aggregation with three granular-part attentions (1, 2, and 4 partitions). $\mathcal{L}_{PCE}$ encourages the attentions from CPA to focus on different sub-areas. In Table 4, employing $\mathcal{L}_{PCE}$ brings 0.3%/0.8% mAP/Rank-1 gains, which is the same as the part-based approach (in the second row in Table 5). Thus, employing $\mathcal{L}_{PCE}$ can only successfully separate the attention parts and lead to aggregation between separate part information. $\mathcal{L}_{PA}$ performs a comparison with the global attention and spatially unifies the amount of information in each attention, which allows separate attentions to focus on the more semantic areas, making all attention parts useful in the aggregation. However, the comparison with the global attention inevitably includes some scores for the background; thus, $\mathcal{L}_{PA}$ alone does not ensure that the attention parts focus on the fully semantic area. To address this problem, learning only the residuals excluding the background scores (*res*) results in a 1.6%/1.8% mAP/Rank-1 performance increment in the last row of Table 4. The improvement proves that CPA is effective.

**Table 4.** Performance of consistent part-attention (CPA) module in MARS [2] under different loss components. **Bold** denotes best performance.

| $\mathcal{L}_{PCE}$ | $\mathcal{L}_{PA}$ | *res* | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|---|
| | | | 85.2 | 89.1 | 86.7 | 97.5 |
| ✓ | | | 85.5 | 89.9 | 96.9 | 97.7 |
| ✓ | ✓ | | 85.6 | 90.1 | 96.9 | 97.9 |
| ✓ | ✓ | ✓ | **86.8** | **90.9** | **97.4** | **98.0** |

For a more detailed analysis of CPA, we provide a few contextual comparisons to visualize how the methods within CPA affect the actual attention map. A comparison of the attention map visualizations for CPA is presented in Figure 7. Using only $\mathcal{L}_{PCE}$ to learn to be classifiable for each attention (2nd row) results in partitions separated by a constant amount, like the horizontal partition (1st row). However, it concurrently generates significant unnecessary background scores. When $\mathcal{L}_{PA}$ is added without the residual technique, although it contains information about the same part with a different situation or identity, certain specific parts still retain unnecessary information, such as the background or occlusion area. This observation indicates that some parts have learned to carry attention scores for irrelevant information from global attention. When the complete CPA framework is applied, the attention remains consistent across all situations, effectively suppressing extraneous information. This shows that CPA eliminates occlusions and misalignments, providing consistent attention regardless of various contexts.

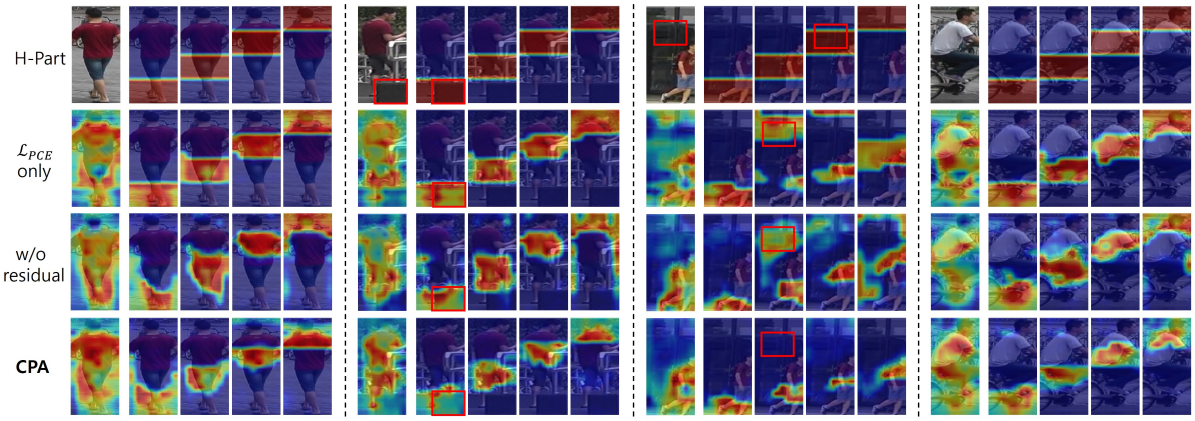


**Figure 7.** Comparison of attention map visualizations for consistent part-attention (CPA) analysis between different situations (e.g., general, occlusion, misaligned, and pose variations). When applying the full CPA, each situation displays a proper and uniformly divided attention map.

**Table 5.** Comparison between part-based method and consistent part attention (CPA) in MARS [2]. **Bold** denotes best performance.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| Baseline | 85.2 | 89.1 | 86.7 | 97.5 |
| Part 4 | 85.5 | 89.9 | 96.9 | 97.7 |
| Part 8 | 85.6 | 90.1 | 96.9 | 97.9 |
| CPA 2 | 86.5 | 90.7 | 97.0 | 97.8 |
| CPA 4 | **86.7** | 91.4 | 97.0 | 97.9 |
| CPA 8 | **86.7** | **91.6** | **97.1** | **98.1** |

#### 4.3.2. Comparison of Part-Based and CPA Models

We conducted experiments to compare the proposed CPA framework with the previous part-based approach. With the same MGA module, we only separated the components of different partitioning methods and the number of parts. The results are presented in

Table 5. Compared with the baseline, the CPA model with 4 partitions achieved a 1.5%/2.3% mAP/Rank-1 increment on MARS. Compared with the part-based approach with the same 4 partitions, the model achieved a 1.2%/1.5% mAP/Rank-1 performance increment. The model achieved the best performance for 8 partitions of CPA with a 1.5%/2.5% mAP/Rank-1 increment compared to the baseline. All CPA partitions achieved more remarkable performance than previous horizontal part-based approaches. This result indicates that the proposed CPA method is more informative than the simple horizontal partition by suppressing unnecessary background information and representing only relevant regions. As shown in Figure 8, depending on the number of parts, each attention part uniformly focuses on the most salient semantic part (e.g., Figure 8c, each attention focuses on a semantic part, such as the head, shirt, pants, and shoe). However, if attention parts are separated too much, each part has very little information, making it difficult to determine the semantic area, as depicted in the third image in Figure 8d. The superiority of the proposed CPA approach over the part-based method is observable in the first and last rows in Figure 7.

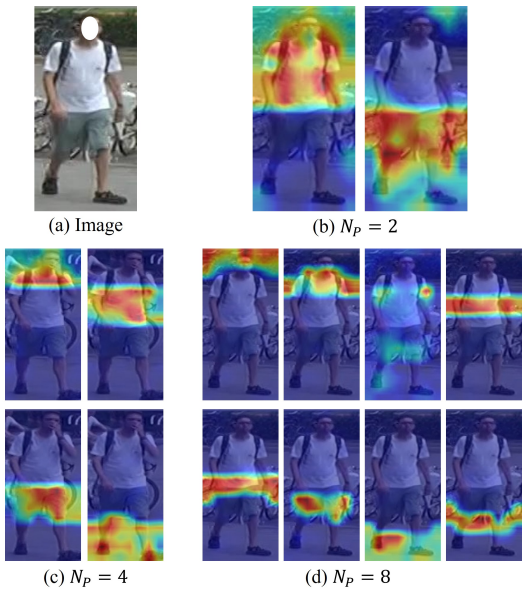


**Figure 8.** Visualization of part attention from the consistent part-attention (CPA) module with a different number of partitions $N_p$.

#### 4.3.3. The Influence of Granularity

To assess the effectiveness of MA-PA and LS-TA as a function of granularity, we compared different combinations of granularity scales in the spatial and temporal dimensions. Table 6 shows the details. First, following [15], we experimented with spatially scaling parts by a factor of 2. The Rank-1 performance increases steadily when more detailed spatial granularities are captured. The scaling factors (i.e., 1, 2, 4, 8) resulted in the highest Rank-1 performance with a 2.8% increment compared to the baseline, showing that the MA-PA approach is effective even with deep granularity. This indicates that the model is stronger in feature extraction when semantic information is diversified. Next, we experimented with different granular factors over temporal cues. We observed that the performance saturates when using the optimal granular in the spatial (i.e., 1, 2, 4) and temporal (i.e., 1, 3, 5) domains, respectively. We achieved a performance increment of 2.0%/2.9% mAP/Rank-1. This indicates that the model supplemented it using useful information when various sampling information was delivered over time.

**Table 6.** Comparison between 'spatial' and 'temporal' scaling factors under different granularity parts in MARS [2]. **Bold** denotes best performance.

| Spatial | Temporal | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|
| - | - | 85.2 | 89.1 | 86.7 | 97.5 |
| 1, 2 | 1 | 86.5 | 90.7 | 97.0 | 97.8 |
| 1, 2, 4 | 1 | **86.7** | 91.1 | **97.5** | **98.0** |
| 1, 2, 4, 8 | 1 | **86.7** | **91.9** | 97.2 | 97.8 |
| 1, 2, 4 | 1, 3 | 86.8 | 90.6 | 97.2 | 97.9 |
| 1, 2, 4 | 1, 3, 5 | **87.2** | **92.0** | 97.3 | 98.1 |
| 1, 2, 4, 8 | 1, 3 | 87.0 | 90.9 | **97.4** | **98.1** |
| 1, 2, 4, 8 | 1, 3, 5 | 86.9 | 91.2 | **97.4** | 98.0 |

#### 4.3.4. Effectiveness of ST-MGA

In ST-MGA, we aggregated the partial features in the spatial and temporal domains. First, the partwise aggregator was used to aggregate across part features, and the scalewise aggregator combines across the granularity scales. We first compared different components of the MGA to validate the partwise and scalewise aggregators. As listed in Table 7, we validate different MGA components on the presence or absence of part-specific aggregators ('part') and scale-specific aggregators ('scale'). A 0.9%/0.5% mAP/Rank-1 performance increment occurred when using only the partwise aggregator, and a 0.7%/0.4% mAP/Rank-1 performance increment occurred when using only the scalewise aggregator. Combining both improved the 0.9%/1.5% mAP/Rank-1 performance, indicating that the proposed combination is more effective than simply averaging the partial features (1st row).

**Table 7.** Performance of ST-MGA in MARS [2] on the presence or absence of partwise aggregator (**'Part'**) and scalewise aggregator (**'Scale'**) in MGA architecture. **Bold** denotes best performance.

| Part | Scale | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|
| | | 86.3 | 90.5 | 97.1 | 97.8 |
| ✓ | | **87.2** | 91.0 | **97.3** | **98.1** |
| | ✓ | 87.0 | 90.9 | **97.3** | **98.1** |
| ✓ | ✓ | **87.2** | **92.0** | **97.3** | 98.0 |

To further validate ST-MGA, we conducted comparative experiments on the influence of each MGA in the spatial (S-MGA) and temporal (T-MGA) domains. Table 8 presents the details. Compared with the baseline, S-MGA achieved a 1.7%/1.8% mAP/Rank-1 increment on MARS. Compared to a simple horizontal partition (P-MGA), S-MGA performed better with a 1.0%/0.4% mAP/Rank-1 increment. Combining T-MGA and P-MGA showed little performance difference compared with T-MGA. The reason may be that T-MGA does not work well for simple horizontal parts due to problems such as temporal misalignment. In contrast, when T-MGA and S-MGA were used together, a 1.1%/1.2% mAP/Rank-1 performance increment occurred compared to only using T-MGA. This result is because S-MGA only deals with semantic information through CPA, so the effect of T-MGA is complementary. To verify which order is better between T-MGA and S-MGA, we made comparisons of the above two different orders. For the Rank-1 metric only, T-MGA $\rightarrow$ S-MGA outperformed by 0.5% compared to using S-MGA $\rightarrow$ T-MGA. As a result, it was verified that the proposed model was effective by exploring knowledge in the spatial and temporal domains, regardless of order, and we used T-MGA $\rightarrow$ S-MGA with relatively high Rank-1 performance as the optimal model.

**Table 8.** Performance of ST-MGA in MARS [2] under different spatial and temporal combinations. **Bold** denotes best performance.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| Baseline | 85.2 | 89.1 | 86.7 | 97.5 |
| P-MGA | 85.9 | 90.5 | 96.5 | 97.6 |
| S-MGA | 86.9 | 90.9 | 97.2 | **98.1** |
| T-MGA | 86.1 | 90.8 | 97.0 | 97.8 |
| T-MGA → P-MGA | 86.0 | 90.9 | 96.9 | 97.8 |
| S-MGA → T-MGA | **87.2** | 91.5 | **97.3** | **98.1** |
| T-MGA → S-MGA | **87.2** | **92.0** | **97.3** | 98.0 |

### *4.4. Comparison and Visualization*

#### 4.4.1. Comparison with the State of the Art

As presented in Table 9, we compared ST-MGA with state-of-the-art methods on the MARS, Duke-V, and LS-VID datasets. For the MARS [2] dataset, ST-MGA reached 87.2% mAP and 92.0% Rank-1 results, which is the best mAP and Rank-5 accuracy, and achieved the second-best Rank-1 results. For the Duke-V [68] dataset, the best mAP and Rank-5 accuracy were achieved at 96.8% and 99.9%, respectively. For the LS-VID [27] dataset, the results were competitive, with the third-best mAP and Rank-1 at 79.3% and 88.5%, respectively. We noticed that ST-MGA outperformed ST-GCN [14] and MGH [15], which use a horizontal part-based approach. Specifically, the proposed method also outperformed TCLNet [18] and BiCnet-TKS [42], which usesimilar diverse attention-based methods, with an improvement of up to 1.4%/2.2% and 1.2%/1.8% mAP/Rank-1 accuracy in MARS, respectively. Further, ST-MGA outperformed several recent models (i.e., SINet [37], CAVIT [38], HMN [40], SGMN [41], and BIC+LGCN [42]). In particular, the proposed method shows higher accuracy than the complex transformer-based method [38], which has recently attracted attention. The above results verify the effectiveness and superiority of ST-MGA in video ReID.

**Table 9.** Quantitative comparison with state-of-the-art methods. **Red** denotes best performance, and **Blue** and **green** denote the second- and third-best performance, respectively.

| Method | MARS | | | Duke-V | | | LS-VID | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAP | Rank1 | Rank5 | mAP | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 |
| M3D [28] | 74.1 | 84.4 | - | - | - | - | 40.1 | 57.7 | - |
| STA [7] | 80.8 | 86.3 | 95.7 | 94.9 | 96.2 | **99.3** | - | - | - |
| GLTR [27] | 78.5 | 87.0 | 95.8 | 93.7 | 96.3 | **99.3** | 44.3 | 63.1 | **77.2** |
| RGSA [29] | 84.0 | 89.4 | - | 95.8 | 97.2 | - | - | - | - |
| FGRA [30] | 81.2 | 87.3 | 96.0 | - | - | - | - | - | - |
| MG-RAFA [26] | 85.8 | 90.0 | 86.7 | - | - | - | - | - | - |
| PhD [31] | 85.8 | 90.0 | 96.7 | - | - | - | - | - | - |
| STGCN [14] | 83.7 | 90.0 | 86.4 | 95.7 | 97.3 | **99.3** | - | - | - |
| MGH [15] | 85.8 | 90.0 | 96.7 | - | - | - | - | - | - |
| TCLNet [18] | 85.8 | 89.8 | - | 96.2 | 96.9 | - | - | - | - |
| AP3D [32] | 85.1 | 90.1 | - | 95.6 | 96.3 | - | - | - | |
| AFA [33] | 82.9 | 90.2 | 96.6 | 95.4 | 97.2 | **99.4** | - | - | - |
| SSN3D [34] | 86.2 | 90.1 | 96.6 | 96.3 | 96.8 | 98.8 | - | - | - |
| BiCnet-TKS [19] | 86.0 | 90.2 | - | 96.1 | 96.3 | - | 75.1 | 84.6 | - |
| STMN [35] | 84.5 | 90.5 | - | 95.9 | 97.0 | | 69.2 | 82.1 | - |
| STRF [36] | 86.1 | 90.3 | - | 96.4 | **97.4** | - | - | - | - |
| SINet [37] | 86.2 | 91.0 | - | - | - | - | **79.6** | 87.4 | - |
| CAVIT [38] | **87.2** | 90.8 | - | - | - | - | 79.2 | **89.2** | - |
| SANet [39] | 86.0 | **91.2** | **97.1** | **96.7** | **97.7** | **99.9** | - | - | - |
| HMN [40] | 82.6 | 88.5 | 96.2 | 96.1 | 96.3 | - | - | - | - |
| SGMN [41] | 85.4 | 90.8 | - | 96.3 | 96.9 | - | - | - | - |
| BIC+LGCN [42] | **86.5** | 91.1 | **97.2** | **96.5** | 97.1 | 98.8 | - | - | - |
| IRNet-V [20] | **87.0** | **92.5** | - | - | - | - | **80.5** | **89.4** | - |
| **ST-MGA** (ours) | **87.2** | **92.0** | **97.3** | **96.8** | **97.6** | **99.9** | **79.3** | **88.5** | **96.1** |

#### 4.4.2. Visualization Analysis

In this section, we visualize some retrieval results from MARS [2] in Figure 9. As can be observed, it is difficult for the baseline model to accurately distinguish people who have a similar appearance or interference when there are an inaccurate detection, occlusion, and uncertain pose. In these cases, the baseline model is not properly focused on the semantic area, and includes unnecessary information about interferences and noises, resulting in relatively low Rank-1 accuracy. The ST-MGA, on the other hand, focused on semantic areas in each part and aggregated them, resulting in high Rank-1 accuracy, even when exposed to relatively difficult and complex interference and noise.

Moreover, we visualize part attention using the proposed CPA module in Figure 10. As observed, the attention parts from CPA focus on similar semantic areas, regardless of the differences between datasets. They each focus on different semantic parts of the face, top, bottom, and shoe and uniformly provide part cues. In occlusion situations, such as in Duke-V, we explored only the correct semantic area, excluding the occluded part. In particular, by focusing on the same semantic part, we can eliminate unnecessary interference in the temporal domain. This allows ST-MGA to contribute to complementary feature extraction by minimizing the interference in both the spatial and temporal domains.

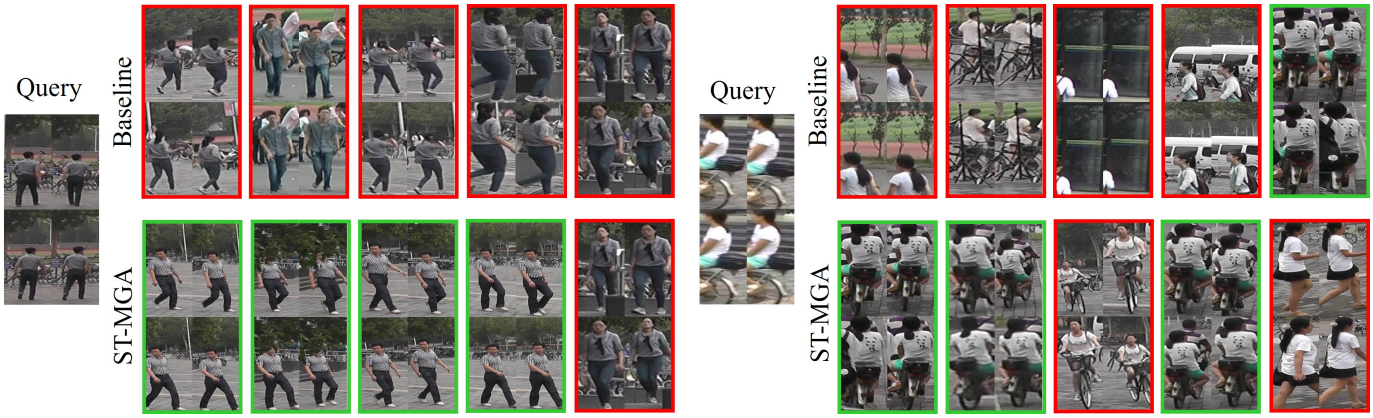


**Figure 9.** The Visualization of ReID retrieval results using the baseline and the proposed ST-MGA on MARS. For each row, the first sequence is the query, while the five sequences in the middle correspond to the Rank-1 to Rank-5 of the baseline model, and the rest are the retrieval results of our ST-MGA. The correct and incorrect matches are marked with green and red bounding boxes, respectively.

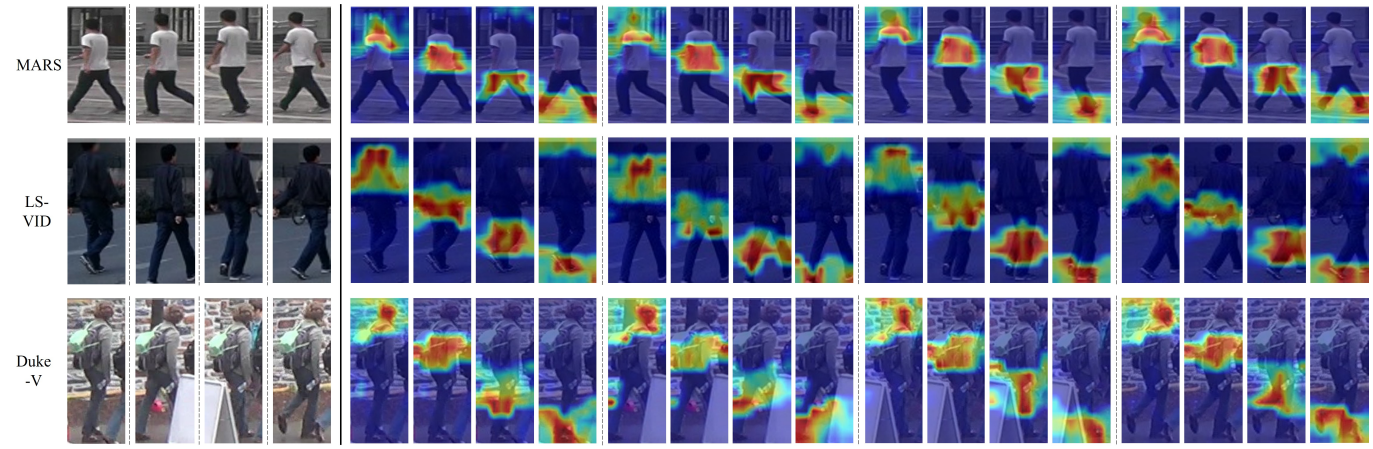


**Figure 10.** Visualization of part attentions from the consistent part-attention (CPA) module on MARS, LS-VID, and Duke-V. Showing the original image and four separated part attentions for each image for different datasets.

## 5. Conclusions

This paper presents the ST-MGA framework, a novel approach for robust video-based person ReID. ST-MGA effectively captures consistent spatiotemporal information to mitigate interference and enhances feature extraction through comprehensive utilization of diverse spatiotemporal granular information. To tackle interference arising from spatiotemporal inconsistency, we introduce the CPA module, which learns from self-information and specific priors. The CPA module efficiently separates attention parts to extract features with spatially uniform amounts and temporally identical semantic information. Additionally, our approach employs multi-granularity feature augmentation to synthesize granular information encompassing semantic attention parts across various scales. Spatially, MA-PA extracts various semantic granular information by synthesizing fine attention without overlapping. Temporally, LS-TA augmented various granular features through various time sampling intervals. Leveraging granular information with different scales, the MGA module effectively utilizes spatiotemporal cues to extract complementary features. Within MGA, we explored the relationships between part/scale information, aggregating them based on relation scores. The resulting aggregated representations enable complementary feature extraction by prioritizing pertinent semantic information while filtering out unnecessary interference or noise. Extensive experiments corroborated the effectiveness and superiority of our ST-MGA, highlighting its potential for advancing video-based person ReID research.

**Author Contributions:** Conceptualization, H.S.L.; Methodology, H.S.L.; Software, H.S.L.; Validation, H.S.L. and M.K.; Formal analysis, M.K. and S.J.; Writing—original draft, H.S.L.; Writing—review & editing, M.K., S.J. and H.B.B.; Visualization, H.B.B.; Supervision, S.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0 02068, Artificial Intelligence Innovation Hub).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Liu, K.; Ma, B.; Zhang, W.; Huang, R. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3810–3818.
2. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VI 14; Springer: Cham, Switzerland, 2016; pp. 868–884.
3. Liu, Z.; Wang, Y.; Li, A. Hierarchical integration of rich features for video-based person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3646–3659. [CrossRef]
4. Kim, M.; Cho, M.; Lee, S. Feature Disentanglement Learning with Switching and Aggregation for Video-based Person Re-Identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1603–1612.
5. Chen, D.; Li, H.; Xiao, T.; Yi, S.; Wang, X. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1169–1178.
6. McLaughlin, N.; Del Rincon, J.M.; Miller, P. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1325–1334.
7. Fu, Y.; Wang, X.; Wei, Y.; Huang, T. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8287–8294.

8. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4733–4742.
9. Liu, Y.; Yan, J.; Ouyang, W. Quality aware network for set to set recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5790–5799.
10. Gao, J.; Nevatia, R. Revisiting temporal modeling for video-based person reid. *arXiv* **2018**, arXiv:1805.02104.
11. Li, S.; Bak, S.; Carr, P.; Wang, X. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 369–378.
12. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.S. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 4913–4922.
13. Liu, Y.; Yuan, Z.; Zhou, W.; Li, H. Spatial and temporal mutual promotion for video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8786–8793.
14. Yang, J.; Zheng, W.S.; Yang, Q.; Chen, Y.C.; Tian, Q. Spatial-temporal graph convolutional network for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3289–3299.
15. Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; Shao, L. Learning multi-granular hypergraphs for video-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2899–2908.
16. Wu, Y.; Bourahla, O.E.F.; Li, X.; Wu, F.; Tian, Q.; Zhou, X. Adaptive graph representation learning for video person re-identification. *IEEE Trans. Image Process.* **2020**, *29*, 8821–8830. [CrossRef] [PubMed]
17. Liu, C.T.; Wu, C.W.; Wang, Y.C.F.; Chien, S.Y. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv* **2019**, arXiv:1908.01683.
18. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Temporal complementary learning for video person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXV 16; Springer: Cham, Switzerland, 2020; pp. 388–405.
19. Hou, R.; Chang, H.; Ma, B.; Huang, R.; Shan, S. 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, Nashville, TN, USA, 20–25 June 2021; pp. 2014–2023.
20. Tao, H.; Duan, Q.; An, J. An Adaptive Interference Removal Framework for Video Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5148–5159. [CrossRef]
21. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
25. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
26. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. 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, Seattle, WA, USA, 13–19 June 2020; pp. 10407–10416.
27. Li, J.; Wang, J.; Tian, Q.; Gao, W.; Zhang, S. Global-local temporal representations for video person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3958–3967.
28. Li, J.; Zhang, S.; Huang, T. Multi-scale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8618–8625.
29. Li, X.; Zhou, W.; Zhou, Y.; Li, H. Relation-guided spatial attention and temporal refinement for video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11434–11441.
30. Chen, Z.; Zhou, Z.; Huang, J.; Zhang, P.; Li, B. Frame-guided region-aligned representation for video person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10591–10598.
31. Zhao, J.; Qi, F.; Ren, G.; Xu, L. Phd learning: Learning with pompeiu-hausdorff distances for video-based vehicle re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2225–2235.

32. Gu, X.; Chang, H.; Ma, B.; Zhang, H.; Chen, X. Appearance-preserving 3d convolution for video-based person re-identification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Cham, Switzerland, 2020; pp. 228–243.
33. Chen, G.; Rao, Y.; Lu, J.; Zhou, J. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VIII 16; Springer: Cham, Switzerland, 2020; pp. 660–676.
34. Jiang, X.; Qiao, Y.; Yan, J.; Li, Q.; Zheng, W.; Chen, D. SSN3D: Self-separated network to align parts for 3D convolution in video person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35; pp. 1691–1699.
35. Eom, C.; Lee, G.; Lee, J.; Ham, B. Video-based person re-identification with spatial and temporal memory networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12036–12045.
36. Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A.K.; Wu, Z. Spatio-temporal representation factorization for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 152–162.
37. Bai, S.; Ma, B.; Chang, H.; Huang, R.; Chen, X. Salient-to-broad transition for video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7339–7348.
38. Wu, J.; He, L.; Liu, W.; Yang, Y.; Lei, Z.; Mei, T.; Li, S.Z. CAViT: Contextual Alignment Vision Transformer for Video Object Re-identification. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 549–566.
39. Bai, S.; Ma, B.; Chang, H.; Huang, R.; Shan, S.; Chen, X. SANet: Statistic attention network for video-based person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3866–3879. [CrossRef]
40. Wang, Z.; He, L.; Tu, X.; Zhao, J.; Gao, X.; Shen, S.; Feng, J. Robust video-based person re-identification by hierarchical mining. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 8179–8191. [CrossRef]
41. Chen, C.; Ye, M.; Qi, M.; Wu, J.; Liu, Y.; Jiang, J. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6100–6112. [CrossRef]
42. Pan, H.; Liu, Q.; Chen, Y.; He, Y.; Zheng, Y.; Zheng, F.; He, Z. Pose-Aided Video-based Person Re-Identification via Recurrent Graph Convolutional Network. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7183–7196. [CrossRef]
43. He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; Hua, X.S. Dense interaction learning for video-based person re-identification supplementary materials. *Identities* **2021**, *1*, 300.
44. Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; Tan, T. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4747–4756.
45. Liao, X.; He, L.; Yang, Z.; Zhang, C. Video-based person re-identification via 3d convolutional networks and non-local attention. In Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Revised Selected Papers, Part VI 14; Springer: Cham, Switzerland, 2019; pp. 620–634.
46. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
47. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
48. Song, G.; Leng, B.; Liu, Y.; Hetang, C.; Cai, S. Region-based quality estimation network for large-scale person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
49. Chen, B.; Deng, W.; Hu, J. Mixed high-order attention network for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 371–381.
50. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 365–381.
51. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3219–3228.
52. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
53. Chen, H.; Zhao, Y.; Wang, S. Person Re-Identification Based on Contour Information Embedding. *Sensors* **2023**, *23*, 774. [CrossRef] [PubMed]
54. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3186–3195.
55. Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. Abd-net: Attentive but diverse person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8351–8361.

56. Kim, M.; Cho, M.; Lee, H.; Cho, S.; Lee, S. Occluded person re-identification via relational adaptive feature correction learning. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2719–2723.
57. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2119–2128.
58. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1179–1188.
59. Si, J.; Zhang, H.; Li, C.G.; Kuen, J.; Kong, X.; Kot, A.C.; Wang, G. Dual attention matching network for context-aware feature sequence based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5363–5372.
60. Liu, J.; Zha, Z.J.; Wu, W.; Zheng, K.; Sun, Q. Spatial-temporal correlation and topology learning for person re-identification in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4370–4379.
61. Tao, H.; Lu, M.; Hu, Z.; Xin, Z.; Wang, J. Attention-aggregated attribute-aware network with redundancy reduction convolution for video-based industrial smoke emission recognition. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7653–7664. [CrossRef]
62. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
63. Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; Wang, D. Pyramid spatial-temporal aggregation for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12026–12035.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
65. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
66. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
67. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
68. Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5177–5186.
69. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Densely semantically aligned person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 667–676.
70. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
71. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.
72. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
73. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 20 February 2024).