# Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification

Yichao Yan[1*], Jie Qin[1*†], Jiaxin Chen[1], Li Liu[1], Fan Zhu[1], Ying Tai[2], and Ling Shao[1]
[1] Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
[2] Tencent YouTu Lab, Shanghai, China
{firstname.lastname}@inceptioniai.org, yingtai@tencent.com

## Abstract

*Video-based person re-identification (re-ID) is an important research topic in computer vision. The key to tackling the challenging task is to exploit both spatial and temporal clues in video sequences. In this work, we propose a novel graph-based framework, namely* ***Multi-Granular Hypergraph*** *(MGH), to pursue better representational capabilities by modeling spatiotemporal dependencies in terms of multiple granularities. Specifically, hypergraphs with different spatial granularities are constructed using various levels of part-based features across the video sequence. In each hypergraph, different temporal granularities are captured by hyperedges that connect a set of graph nodes (i.e., part-based features) across different temporal ranges. Two critical issues (misalignment and occlusion) are explicitly addressed by the proposed hypergraph propagation and feature aggregation schemes. Finally, we further enhance the overall video representation by learning more diversified graph-level representations of multiple granularities based on mutual information minimization. Extensive experiments on three widely-adopted benchmarks clearly demonstrate the effectiveness of the proposed framework. Notably,* ***90.0%*** *top-1 accuracy on* ***MARS*** *is achieved using MGH, outperforming the state-of-the-arts. Code will be released at* https://github.com/daodaofr/hypergraph_reid.

## 1. Introduction

Person re-identification (re-ID) aims at associating individuals across non-overlapping cameras, with great potential in surveillance-related applications. As such, significant efforts have been made in the past few years to address the challenging task. In parallel with the prevalence of image-based person re-ID, person re-ID based on video sequences has also recently emerged. This is because the richer information in videos can be utilized to reduce visual ambiguities, especially for people sharing similar appearances. The key to solving video-based person re-ID is to concurrently exploit spatial and temporal clues within video sequences. In this sense, this work aims to shed light on two important clues (see Figure 1) for tackling video-based person re-ID.

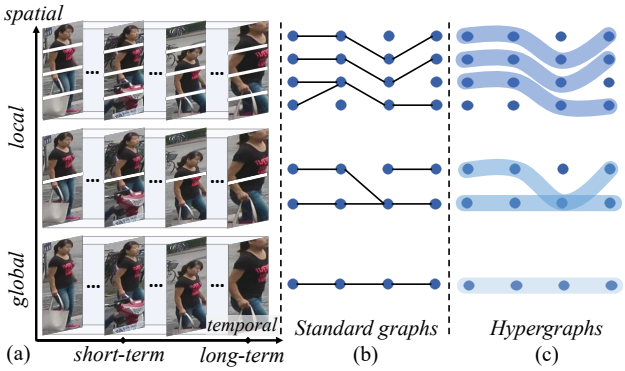


Figure 1: (a) Illustration of multi-granular spatial and temporal clues, which provide important insights into addressing the challenges of misalignment and occlusion in video-based person re-ID. (b) Standard graphs can only model the dependency between node pairs, lacking the capability of modeling long-term temporal dependency. (c) Hypergraphs can model both short-term and long-term dependencies by associating multiple nodes within a single hyperedge.

1) **Multi-granularity of spatial clues.** As the structural information of the human body is beneficial for person identification, part-based models [49, 30, 47] have generally achieved promising performance in person re-ID. Compared with fixed partitions, multi-granular part-based models [51, 64] have further enhanced the performance by dividing the human body into multiple granularities. In video-based re-ID, the multi-granular spatial clues are particularly important since different levels of granularities capture discrepancies between different partitions, thus addressing the spatial *misalignment* issue due to inaccurate detection in the video sequence (see Figure 1(a)). However, in this way, misalignment can only be solved implicitly to some extent. To better tackle spatial misalignment, we need to explicitly align different body parts across the whole sequence in

*indicates equal contributions; † indicates corresponding author

order to achieve more robust re-ID performance. It is therefore highly desirable to develop re-ID within a framework which systematically captures correlations between different body partitions, while the same time being able to exploit multiple spatial granularities.

2) **Multi-granularity of temporal clues.** Temporal clues have been extensively studied by previous video-based re-ID models. Short-term dynamics can be represented by extracting additional optical flow features [9], while long-term temporal features can be obtained by utilizing 3D CNNs [32] or temporal feature aggregators [39] (*e.g.*, Recurrent Neural Networks, RNNs). However, short- and long-term temporal clues have different functionalities in discriminative feature learning. For example, in Figure 1, there is a partial occlusion w.r.t. the short-term temporal clues; the long-term temporal clues can help reduce its impact. However, only a few works [31, 32] address this issue. It is thus highly important to design a model that can capture multi-granular temporal clues.

To explicitly fulfill the above goal, we propose a novel graph-based framework, named **Multi-Granular Hypergraph (MGH)**, which simultaneously exploits spatial and temporal clues for video-based person re-ID. As shown in Figure 2, we construct a set of hypergraphs to model multiple granularities in a video sequence, with graph nodes representing global or part-level features. Each hypergraph models a specific spatial granularity, while each hyperedge, connecting multiple nodes within a particular temporal range, captures a specific temporal granularity. Node-level features are propagated to form graph-level representations for all hypergraphs, which are aggregated in the final video representation to achieve robust person re-ID.

Our MGH method has three main advantages. *First*, it seamlessly unifies the learning of spatial and temporal clues into a joint framework, where spatial clues are captured by different hypergraphs, and short- and long-term temporal clues are mined with message propagation through different hyperedges. *Second*, compared with standard graphs which only model correlations between pairs of nodes (see Figure 1(b)), hypergraphs can model high-order dependencies among multiple nodes (see Figure 1(c)). As a result, *misalignment* can be explicitly solved by associating different nodes with their nearest neighbors using hyperedges; meanwhile, *occlusions* can be addressed by modeling multi-granular temporal dependencies with hyperedges across different temporal ranges. *Third*, node-level features can benefit from the spatial and temporal information in the sequence by means of HyperGraph Neural Networks (HGNNs) [14], which greatly facilitate information propagation through hyperedges. Our main contributions include:

- We formulate video-based person re-ID as a hypergraph learning task, yielding robust representations based on node propagation and feature aggregation.
- To capture multi-granular clues, we design a novel HGNN architecture (*i.e.*, MGH) to simultaneously exploit spatial and temporal dependencies in videos.
- The diversity of graph representations corresponding to different spatial granularities is preserved and enhanced by employing an intuitive loss based on mutual information minimization.
- MGH achieves promising results on three widely-used re-ID benchmarks. Notably, MGH obtains **90.0%** top-1 accuracy on **MARS**, one of the largest video re-ID datasets, outperforming the state-of-the-art models.

## 2. Related Work

**Person Re-identification.** Existing works on person re-ID mainly focus on two sub-tasks, *i.e.*, image-based [16, 66, 67, 7, 2] and video-based [13, 68] person re-ID. Here, we briefly review some closely related works for video-based re-ID. Early methods tend to employ hand-crafted spatiotemporal features, such as HOG3D [28] and SIFT3D [42]. Other methods try to extract more discriminative descriptors[26, 35] or design more effective ranking algorithms [52, 61, 3]. Recently, various deep learning models have been proposed and have shown superior performance compared with hand-crafted features. Some works [65, 33, 57, 46, 34, 62, 36] leverage the powerful learning capability of Convolutional Neural Networks (CNNs) and perform straightforward spatial/temporal pooling on video sequences to generate global representations. However, simply pooling the features may lead to a significant loss of discriminative information. Other methods [39, 60, 70, 37, 5] adopt RNNs and attention mechanisms for more robust temporal feature fusion. However these methods neglect the importance of spatial clues. Another class of methods [9, 39] resort to using additional information on optical flow, and adopt a two-stream structure [44] for discriminative feature learning. However, optical flow only represents local dynamics of adjacent frames, which may introduce noise due to spatial misalignment. 3D CNNs [24, 48] have also been applied to address video-based person re-ID [32]. Despite their promising performance, these networks are computationally expensive and difficult to optimize. In this work, we explicitly explore the multi-granular nature of both spatial and temporal features, yielding more robust representations for video-based re-ID.

**Hypergraph Learning.** Graphs are typically leveraged to model relationships between different nodes. Depending on the type of data the nodes represent, graphs have been explored in many computer vision tasks, such as action recognition [55], image classification [8], and person re-ID [6, 43]. Recently, neural networks have been extensively studied for graph learning, leading to the widespread usage of Graph Neural Networks (GNNs) [41, 10, 12, 29, 53]. How-

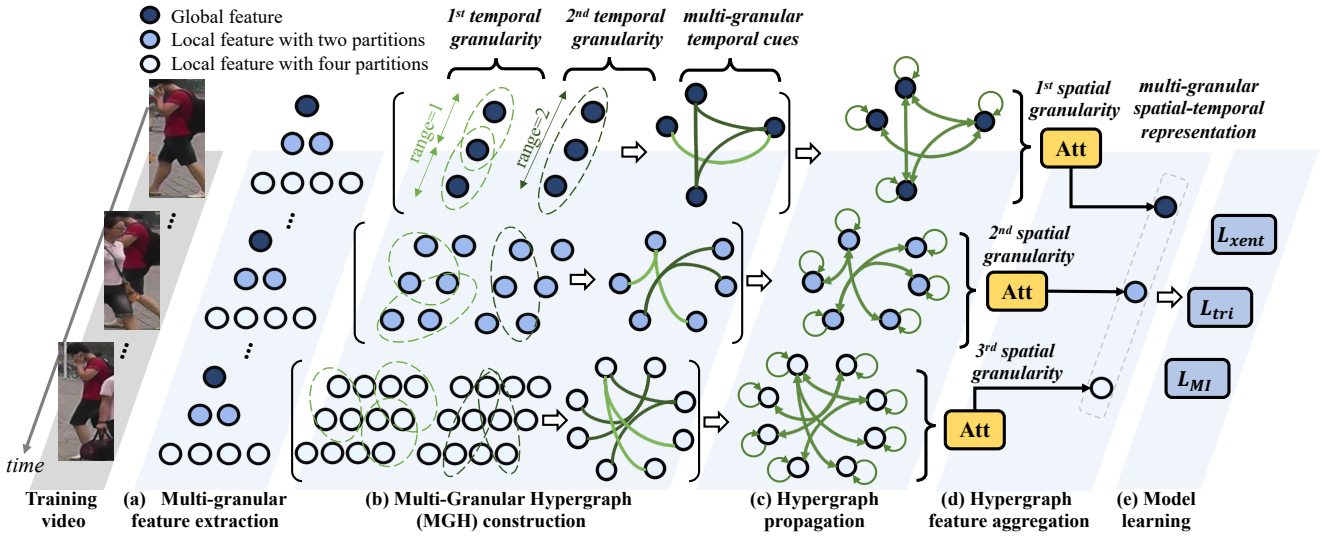


Figure 2: Detailed architecture of the proposed multi-granular hypergraph learning framework for video-based person re-ID. For better visualization, we only illustrate the first three spatial granularities and the first two temporal granularities.

ever, conventional graphs can only model pairwise relationships, which prevents their scalability to data with more complex structures. To this end, hypergraph [69] was introduced to model higher-order relationships between objects of interest, and has been applied to video segmentation [22] and image retrieval [23]. In a similar spirit to GNNs, HyperGraph Neural Networks (HGNNs) [14, 59, 4, 25] have recently been proposed to model correlations in hypergraphs using deep neural networks. Inspired by HGNNs, this work derives a hypergraph that explicitly models spatiotemporal dependency in a video sequence. More importantly, multiple spatial and temporal granularities are exploited simultaneously in our hypergraph learning framework. As a result, the final global representation exhibits strong discriminability for robust video-based re-ID.

## 3. Multi-Granular Hypergraph Learning

Although deep learning and temporal modeling approaches have greatly improved the performance of video-based person re-ID, it is still difficult to achieve satisfactory results because of occlusion, misalignment, background clutter, and viewpoint changes. To further improve the discriminability of feature representations, this paper aims to explicitly explore the multi-granular nature of spatial and temporal features. To this end, we design a Multi-Granular Hypergraph (MGH) learning framework, which models the high-order correlations between spatial and temporal clues with a hypergraph neural network. The details of the proposed framework are elaborated as follows.

### 3.1. Multi-Granular Feature Extraction

Recent studies [51, 64] have demonstrated that multi-granular spatial features have the advantage of generating more discriminative representations for human bodies. Inspired by this, we extract multi-granular features for individuals. Specifically, given an image sequence $\mathbf{I} = \{I_1, I_2, ..., I_T\}$ containing $T$ images, we use a backbone CNN model to extract individual feature maps

$$\mathbf{F}_i = \text{CNN}(I_i), i = 1, ..., T, \tag{1}$$

where $\mathbf{F}_i$ is a 3D tensor with dimensions $C \times H \times W$. $C$ is the channel size, and $H$ and $W$ are the height and width of the feature map, respectively. We then hierarchically divide the feature maps into $p \in \{1, 2, 4, 8\}$ horizontal parts w.r.t. different levels of granularities, and perform average pooling on the divided feature maps to construct a part-level feature vector. For each granularity, the whole sequence generates $N_p = T \times p$ part-level features, which we denote as $\mathbf{h}^0 = \{\mathbf{h}^0_1, \mathbf{h}^0_2, ..., \mathbf{h}^0_{N_p}\}$. For example, in each video frame, the first granularity contains a single global vector, while the second and third granularities contain two and four part-level features, respectively, as shown in Figure 2(a).

### 3.2. Multi-Granular Hypergraph

After extracting the initial global or part-based features of each individual, *i.e.* initial node features, the next step is to update the node features by learning correlations among different nodes. To generate robust representations, it is necessary to take into account both the spatial and temporal correlations of individual features. Inspired by the recent success of HGNNs [14, 25], we propose a novel hypergraph neural network for spatiotemporal feature learning. To explore the spatial and temporal dependencies within a sequence, HGNNs allow nodes to communicate with their neighbors through message passing within the graph. More importantly, compared to standard graph mod-

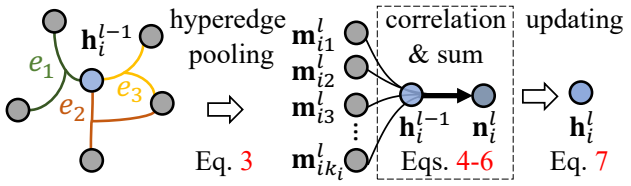


Figure 3: Illustration of node feature propagation.

els, hypergraphs can model the high-order dependency involving multiple nodes, which is more flexible and suitable for modeling the multi-granular correlations in a sequence.

**Hypergraph Construction.** We propose to capture the spatial and temporal dependencies by constructing a set of hypergraphs $\mathcal{G} = \{\mathcal{G}_p\}_{p\in\{1,2,4,8\}}$, where each hypergraph corresponds to a specific spatial granularity. Concretely, $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$ consists of $N_p$ vertices $\mathcal{V}_p$ and a set of hyperedges $\mathcal{E}_p$. Here, we utilize $v_i \in \mathcal{V}_p$, where $i \in \{1, ..., N_p\}$ to denote the $i$-th graph node. We define a set of hyperedges to model short- to long-term correlations in the hypergraph. To learn short-term correlations, a hyperedge only connects temporally adjacent features. Mid- and long-range correlations are modeled by hyperedges connecting features of different temporal lengths. Specifically, for each graph node $v_i$, we find its $K$ nearest neighbors within specific temporal ranges, according to the feature affinities between nodes. Then we utilize a hyperedge to connect these $K$+1 nodes, as shown in Figure 2(b). Mathematically,

$$e_{it} = \{v_i, \forall v_j \in \mathcal{N}_K(v_i)\}, s.t. \ |v_i - v_j| \leq T_t, \quad (2)$$

where $\mathcal{N}_K$ is the neighborhood set containing the top-$K$ neighbors, $|*|$ denotes the temporal distance between the vertices in the sequence, and $T_t$ is the threshold of temporal range. In our framework, we adopt three thresholds (*i.e.*, $T_1, T_2, T_3$) to model short-term, mid-term and long-term dependencies, respectively.

**Hypergraph Propagation.** Based on the hypergraphs, we design a hypergraph neural network to propagate graph information and update node features, as illustrated in Figure 3. Given a node $v_i$, we use $Adj(v_i) = \{e_1, e_2, ..., e_{k_i}\}$ to denote all the hyperedges that include this node. These hyperedges contain the nodes that have the highest correlations with $v_i$. Then an aggregation operation is defined on the hyperedges to capture feature correlations. Specifically, we average all node features in a hyperedge, except for $v_i$, as the hyperedge feature w.r.t. this node:

$$\mathbf{m}_{ik}^l = \sum_{j\neq i}^{v_j\in e_k} \mathbf{h}_j^{l-1}, \ \forall e_k \in Adj(v_i), \quad (3)$$

where $\mathbf{h}_j^{l-1}$ denotes the node feature of $v_j$ in layer $l-1$ of the HGNN. We then calculate the importance of each hyperedge by measuring the correlation between node features

**Algorithm 1** Hypergraph Propagation

**Input:** Input sequence $\mathbf{I} = \{I_1, I_2, ..., I_T\}$
**Output:** Hypergraph feature $O_p$
1: Extract and pool features by Eq. 1: $\mathbf{h}^0 \leftarrow I$
2: Build the hyperedges by Eq. 2
3: **for** $l \leftarrow 1, ..., L$ **do**
4: Pooling hyperedge features by Eq. 3: $\mathbf{m}_{ik}^l \leftarrow \mathbf{h}^{l-1}$
5: Calculate feature correlations and aggregate hyperedge message by Eqs. 4- 6: $\mathbf{n}_i^l \leftarrow \mathbf{m}_{ik}^l$
6: Updating node features by Eq. 7: $\mathbf{h}_i^l \leftarrow \{\mathbf{h}_i^{l-1}, \mathbf{n}_i^l\}$
7: **end for**
8: $\mathbf{O}_p \leftarrow \{\mathbf{h}_i^L\}$.

and hyperedge features:

$$z_{ik} = \phi(\mathbf{h}_i^{l-1}, \mathbf{m}_{ik}^l), \quad (4)$$

where $\phi$ measures the similarity between features (we employ cosine similarity in our framework). We then utilize the Softmax function to normalize the importance weights and aggregate the hyperedge messages as follows:

$$\gamma_{ik} = \frac{\exp(z_{ik})}{\sum_j \exp(z_{ij})}, \quad (5)$$

$$\mathbf{n}_i^l = \sum_k \gamma_{ik} \mathbf{m}_{ik}^l. \quad (6)$$

After obtaining the hypergraph messages, the node features are updated in a fully connected layer by concatenating the previous node features and hyperedge message:

$$\mathbf{h}_i^l = \sigma(\mathbf{W}^l[\mathbf{h}_i^{l-1}, \mathbf{n}_i^l]), \quad (7)$$

where $\mathbf{W}^l$ is a weight matrix and $\sigma$ is an activation function. The above feature updating steps are repeated for $L$ rounds and we obtain a set of output node features $\mathbf{O}_p = \{\mathbf{h}_i^L\}, \forall v_i \in \mathcal{V}_p$. We summarize the propagation process of the hypergraph in Algorithm 1.

**Attentive Hypergraph Feature Aggregation.** After obtaining the final updated node features of each hypergraph w.r.t. each spatial granularity, we further need to aggregate node/part-level features into graph/video-level representations for each hypergraph. When deriving aggregation schemes, we should take into account that, within a hypergraph, different nodes are of varying importance. For instance, the occluded parts or backgrounds are less important than the human body parts. It is therefore necessary to develop a specific attention mechanism [1, 38] to address this. As shown in Figure 4, we propose an attention module which generates the node-level attention for each hypergraph, in order to select the most discriminative part-level features. For each hypergraph, we calculate the node attention $\boldsymbol{\alpha}_p = \{\alpha_1, ..., \alpha_{N_p}\}$ as follows:

$$u_i = \mathbf{W}_u \mathbf{h}_i^L, \quad (8)$$

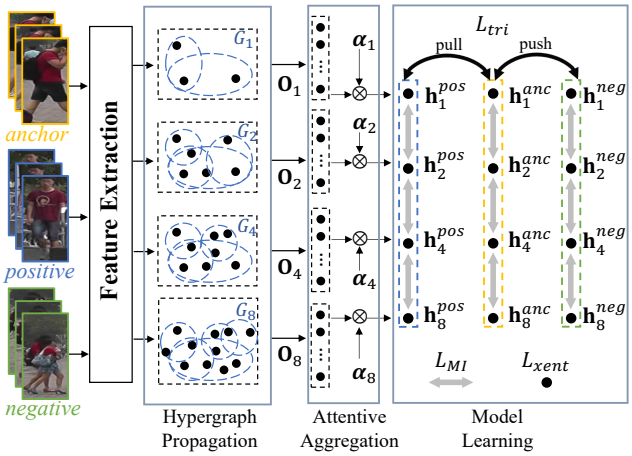

Figure 4: Illustration of attentive node feature aggregation and model learning modules.

$$\alpha_i = \frac{\exp(u_i)}{\sum_j \exp(u_j)}, \tag{9}$$

where $\mathbf{W}_u$ is the weight matrix. The hypergraph features are then calculated as a weighted sum of the node features:

$$\mathbf{h}_p = \sum_{v_i \in \mathcal{V}_p} \alpha_i \mathbf{h}_i^L. \tag{10}$$

### 3.3. Model Learning

To optimize the framework, we adopt the cross entropy loss and triplet loss to jointly supervise the training procedure. The cross entropy loss is formulated as follows:

$$L_{xent} = -\sum_{i=1}^{N} \log \frac{\exp(\mathbf{W}_{y_i}\mathbf{h}_p^i + b_{y_i})}{\sum_{k=1}^{C} \exp(\mathbf{W}_k \mathbf{h}_p^i + b_k)}, \tag{11}$$

where $y_i$ is the label of feature $\mathbf{h}_p^i$, $N$ is the mini-batch size, and $C$ is the number of classes in the training set. Given a triplet consisting of the anchor, positive, and negative features, *i.e.*, $\{\mathbf{h}_p^{anc}, \mathbf{h}_p^{pos}, \mathbf{h}_p^{neg}\}_{\{p=1,2,4,8\}}$, the hard triplet loss is calculated as follows:

$$\begin{aligned} L_{tri} = -\sum_{anc=1}^{N} [m + \max_{pos=1\ldots N} \|\mathbf{h}_p^{anc} - \mathbf{h}_p^{pos}\|_2 \\ - \min_{neg=1\ldots N} \|\mathbf{h}_p^{anc} - \mathbf{h}_p^{neg}\|_2]_+, \end{aligned} \tag{12}$$

where $m$ denotes the margin.

After training the model based on the two loss terms above, each hypergraph will output discriminative graph-level features. The last step is to aggregate graph features w.r.t. different spatial granularities to form the final video representation. In practice, we find that directly pooling graph-level features may lead to significant information loss as each hypergraph captures the unique characteristic of the corresponding granularity. Therefore, we should maintain the diversity of different levels of graph features. Inspired by the information theory, we attempt to fulfill this goal by *mutual information minimization*. Specifically, we adopt an additional loss that reduces the mutual information between features from different hypergraphs, thus increasing the discriminability of the final video representation by concatenating all the features. Here we denote $\mathbf{H}_p = \{\mathbf{h}_p^i\}_{i=1}^{N_c}$ as the graph-level features with $p$ spatial partitions, where $N_c$ is the number of tracklets in the training set. Following [20], we define the mutual information loss as:

$$L_{MI} = \sum_{p,q \in \{1,2,4,8\}}^{p \neq q} \mathcal{I}(\mathbf{H}_p, \mathbf{H}_q), \tag{13}$$

where $\mathcal{I}$ measures the mutual information between different hypergraph features.

Finally, as shown in Figure 4, the overall loss function is a combination of the above three terms:

$$L_{all} = L_{xent} + L_{tri} + L_{MI}. \tag{14}$$

## 4. Experimental Results

We evaluate MGH on three benchmark datasets, *i.e.*, MARS [65], iLIDS-VID [52], and PRID-2011 [19]. We first conduct a comprehensive ablation study to verify the contribution of each component of our model, and then compare our model with recent state-of-the-art approaches.

### 4.1. Experimental Setup

**Datasets.** **MARS** [65] is one of the largest public datasets for video-based person re-ID, which consists of 1,261 pedestrians captured by six cameras, and each individual appears in at least two cameras. Meanwhile, each identity has 13.2 tracklets on average. **iLIDS-VID** [52] contains 600 image sequences of 300 people from two non-overlapping camera views in an airport arrival hall. The frame lengths in each sequence vary from 23 to 192, with an average length of 73. **PRID-2011** [19] was collected in an uncrowded outdoor environment with a relatively clean background. This dataset includes 749 people from two camera views, but only the first 200 are captured by both cameras. The length of sequence varies from 5 to 675, with an average of 100. Following previous practice [52], we only utilize the sequence pairs with more than 21 frames.

**Evaluation Protocols.** In terms of MARS, we use the predefined training/test split, *i.e.*, 8,298 sequences of 625 people are used for training, and 12,180 sequences of 636 people are used for testing. As for iLIDS-VID and PRID-2011, we follow the standard evaluation protocol [52]. People are randomly split into two subsets with equal size as training and test sets, and the performance is reported as the average results of ten trials. For all the datasets, we use

| Methods | MARS | | iLIDS | PRID |
|---|---|---|---|---|
| | mAP | top-1 | top-1 | top-1 |
| baseline | 78.3 | 85.7 | 79.7 | 88.6 |
| HGNN | 84.1 | 88.7 | 84.3 | 93.8 |
| + Att | 84.8 | 89.2 | 85.1 | 94.2 |
| + $L_{MI}$ | **85.8** | **90.0** | **85.6** | **94.8** |

Table 1: Component analysis of MGH. 'Att' denotes node-level attention, '$L_{MI}$' denotes mutual information loss.

Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) to measure the performance.

**Implementation Details.** We employ ResNet50 [18] pretrained on ImageNet [11] as the backbone. Following [21], we insert non-local blocks [54] into the network and we resize the input images to 256×128. For each batch, we randomly sample 32 sub-sequences from 8 persons. In practice, we set the sub-sequence length $T = 8$, the hypergraph layer $L = 2$, and the number of neighbors $K = 3$. The default spatial partitions are $(1, 2, 4, 8)$, and the temporal thresholds are $(1, 3, 5)$. The influences of these hyperparameters are analyzed in Section 4.3. We adopt the Adam [27] optimizer with weight decay 0.0005. The initial learning rate is set to 0.0003 and is reduced by a factor of 10 every 100 epochs, with the training stage terminating at the 300-th epoch. We concatenate the graph-level features and use the cosine similarity as the distance metric for matching the final video representations. All the experiments are implemented in PyTorch [40], with a Tesla V100 GPU.

### 4.2. Model Component Analysis

We evaluate the contribution of each component and report the results in Table 1. For the baseline in the first row, the backbone network is trained with cross entropy and triplet losses. The results are obtained by performing average pooling on frame-level features across the whole video sequence. Through multi-granular spatial and temporal dependency learning with an HGNN, we observe that the top-1 accuracy increases by 3% on MARS, and by 5% on iLIDS-VID and PRID-2011, as shown in the 2nd row of Table 1. We then insert the node attention module for graph-level feature aggregation, the attention modules further improve the top-1 accuracy by 0.5%-1%, respectively. Finally, by further incorporating the mutual information loss, the top-1 accuracy and mAP are respectively improved from 85.7% and 78.3% to 90.0% and 85.8% on MARS. As for iLIDS-VID and PRID-2011, the improvement of top-1 accuracy is more than 5%. In summary, the major improvement of the framework comes from our hypergraph learning mechanism, as it captures the dependencies of both spatial and temporal clues. The attention module and the mutual information loss bring additional improvements by learning more discriminative graph and video-level features.

| Spatial | Temporal | mAP | top-1 |
|---|---|---|---|
| 1 | 1 | 82.3 | 87.5 |
| 1 | 3 | 82.8 | 87.7 |
| 1 | 5 | 82.9 | 87.7 |
| 1 | 1,3 | 83.2 | 87.9 |
| 1 | 1,3,5 | 83.3 | 88.1 |
| 1,2 | 1,3,5 | 84.6 | 89.3 |
| 1,2,4 | 1,3,5 | 85.5 | 89.8 |
| 1,2,4,8 | 1,3,5 | **85.8** | **90.0** |
| 1,2,4,8 | 1,3,5,7 | 85.7 | **90.0** |

Table 2: Performance of MGH on MARS under different granularities of spatial and temporal clues. 'Spatial' denotes the number of human body partitions, and 'Temporal' denotes the temporal range for dependency calculation.

### 4.3. Model Sensitivity Analysis

**Multi-Granularity.** The key motivation of this work is to make full use of the multi-granular spatial and temporal clues in a video sequence to learn better representations. Here, we conduct detailed experiments to evaluate the effectiveness of multi-granular representations, the results of which are illustrated in Table 2. We test different combinations of spatial and temporal granularities. Specifically, when the temporal range is equal to one, only adjacent nodes are connected. We observe that the performance increases steadily when more detailed spatial/temporal granularities are captured. We also find that the performance saturates when using four spatial granularities (*i.e.*, 1, 2, 4, 8) and three temporal granularities (*i.e.*, 1, 3, 5).

**Node Propagation Scheme.** In MGH, we aggregate the hyperedge features by calculating the correlations with the target node, as depicted by Eqs. 4-6. Here, we compare our correlation aggregator with several alternative aggregation schemes. It is worth noting that pooling-based aggregators and LSTM have also been widely utilized for feature aggregation in GNNs [17]. However, as shown in Figure 5(a), they ("max", "avg" and "LSTM") are less effective than the correlation aggregator since they neglect the dependency of hyperedge features w.r.t. the target node. Besides, graph propagation often adopts attention mechanisms [50], which typically concatenate the inputs and utilize a fully connected layer to generate the attention weights. We observe that such an attention scheme achieves comparable performance with our correlation aggregator, but it requires more computational overhead. Overall, these results demonstrate the effectiveness of the correlation aggregator.

**Number of Hypergraph Layers $L$.** We evaluate the influence of different numbers of HGNN layers. From Figure 5(b), we can see that the proposed framework is not sensitive to different numbers of layers. Specifically, a two-layer network achieves slightly better performance than other settings. This is because a one-layer HGNN has

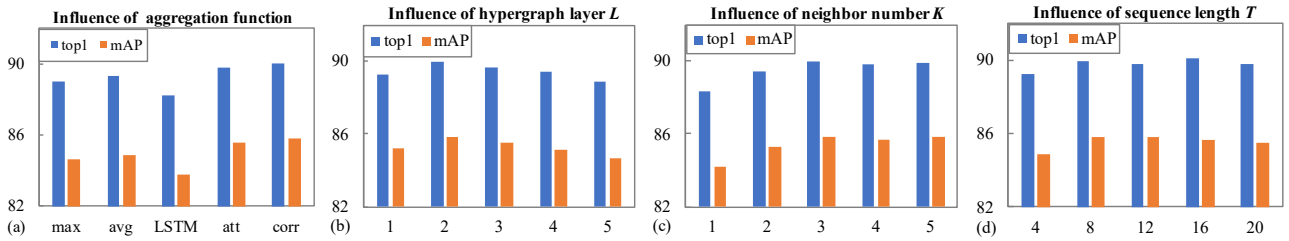


Figure 5: Results on MARS illustrating the influence of different hyperparameter. (a) aggregation function; (b) hypergraph layer $L$; (c) neighbor number $K$; (d) sequence length $T$. Zoom in for best visualization.

| Methods | Source | MARS | | | | iLIDS-VID | | | PRID-2011 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | top-1 | top-5 | top-20 | top-1 | top-5 | top-20 | top-1 | top-5 | top-20 |
| CNN+XQDA [65] | ECCV16 | 47.6 | 65.3 | 82.0 | 89.0 | 53.0 | 81.4 | 95.1 | 77.3 | 93.5 | 99.3 |
| SeeForest [70] | CVPR17 | 50.7 | 70.6 | 90.0 | 97.6 | 55.2 | 86.5 | 97.0 | 79.4 | 94.4 | 99.3 |
| ASTPN [58] | ICCV17 | - | 44 | 70 | 81 | 62 | 86 | 98 | 77 | 95 | 99 |
| STAN [34] | CVPR18 | 65.8 | 82.3 | - | - | 80.2 | - | - | 93.2 | - | - |
| ETAP-Net [57] | CVPR18 | 67.4 | 80.8 | 92.1 | 96.1 | - | - | - | - | - | - |
| Snippet [5] | CVPR18 | 76.1 | 86.3 | 94.7 | 98.2 | 85.4 | 96.7 | 99.5 | 93.0 | 99.3 | 100 |
| STA [15] | AAAI19 | 80.8 | 86.3 | 95.7 | - | - | - | - | - | - | - |
| ADFD [63] | CVPR19 | 78.2 | 87.0 | 95.4 | **98.7** | **86.3** | **97.4** | **99.7** | 93.9 | **99.5** | 100 |
| VRSTC [21] | CVPR19 | **82.3** | **88.5** | **96.5** | 97.4 | 83.4 | 95.5 | 99.5 | - | - | - |
| COSAM [45] | ICCV19 | 79.9 | 84.9 | 95.5 | 97.9 | 79.6 | 95.3 | - | - | - | - |
| GLTR [31] | ICCV19 | 78.5 | 87.0 | 95.8 | **98.2** | **86.0** | **98.0** | - | **95.5** | **100** | 100 |
| AdaptiveGraph [56] | arXiv19 | **81.9** | **89.5** | **96.6** | 97.8 | 84.5 | 96.7 | 99.5 | **94.6** | 99.1 | 100 |
| **MGH** | - | **85.8** | **90.0** | **96.7** | **98.5** | **85.6** | **97.1** | **99.5** | **94.8** | **99.3** | 100 |

Table 3: Comparison with the state-of-the-art video-based person re-id methods. The three best scores are indicated in **red**, **blue** and **green**, respectively.

insufficient representational capability, whilst multi-layer HGNNs contains too many parameters that bring difficulties to the training step. Therefore, we employ a two-layer HGNN in our framework.

**Number of Neighbors $K$.** This hyperparameter controls the number of nodes within a hyperedge. Specifically, when $K = 1$, an edge only connects two graph nodes and the hypergraph degrades into a standard graph. As shown in Figure 5(c), in the beginning, the performance increases as $K$ becomes larger, since more context information is included in the hyperedge. However, the performance becomes saturated when $K > 3$. These results validate the effectiveness of employing a hypergraph rather than a standard graph.

**Sequence Length $T$.** Last but not least, we train and test the framework with various sequence lengths $T$, and the results are illustrated in Figure 5(d). Overall, the proposed framework is robust to variations in $T$. We also find that longer sequences generate slightly better performance, since the model can capture wider ranges of temporal dependencies. Meanwhile, employing longer sequences increases the model complexity. Overall, $T = 8$ gives the best trade-off between performance and complexity.

In summary, the proposed MHG is not sensitive to most hyperparameters in the framework. The granularity of the spatial and temporal clues plays a key role in the overall performance, which aligns well with the motivation of the proposed framework.

### 4.4. Comparisons with State-of-the-Arts

In this section, we compare the proposed MGH with the state-of-the-art methods on three video-based person re-id benchmarks. The results are reported in Table 3.

On MARS, our approach achieves 90% top-1 accuracy, outperforming all previous methods. More remarkably, the proposed model achieves 85.8% in mAP without re-ranking, showing significant improvement (*i.e.*, 3.5% higher) over the current best state-of-the-art method. We notice that the recently proposed AdaptiveGraph [56] also employs GNNs to address the video-based person re-ID task. Our model has two advantages. First, AdaptiveGraph requires additional pose information to build the graph, while our MGH is dynamically constructed based on feature affinities, which makes the proposed model more flexible. Second, AdaptiveGraph only considers the correlation between adjacent frames, neglecting the long-term temporal dependency. In the proposed hypergraph, dependencies in varying temporal ranges are modeled by different hyperedges, yielding more robust representations. It is worth noting that VRSTC [21] also achieves promising results on MARS; however, it is a two-stage model, *i.e.*, VRSTC


Figure 6: Visualization of re-ID results using the baseline model and the proposed MGH model. The first sequence is the query, whilst the rest are the Rank-1 to Rank-4 (from left to right) retrieved results. The green and red bounding boxes denote correct and incorrect matches, respectively.

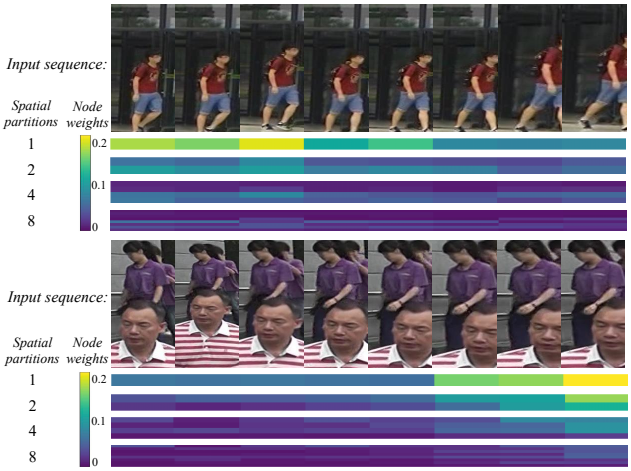


Figure 7: Visualization of node attention weights. The node attention wights are visualized below the input sequences.

first locates the occluded regions and completes the regions with a generative model, and then utilizes the non-occluded frames for re-ID. In contrast, the proposed MGH does not need such pre-processing and can be learned end-to-end.

In terms of iLIDS-VID and PRID, since they only contain a single correct match in the gallery set, we only report the cumulative re-ID accuracy. Overall, the proposed MGH achieves competitive results compared with other state-of-the-art methods. Specifically, MGH outperforms several recent models (*i.e.*, SeeForest [70], ASTPN [58], STAN [34], ETAP-Net [57], Snippet [5] with optical flow inputs, STA [15] and COSAM [45]) in terms of all the evaluation metrics on the two datasets. We note that ADFD [63] obtains strong performance on iLIDS-VID and PRID-2011. This may be because ADFD employs external attribute labels to learn disentangled feature representations, and such additional information is more effective on small-scale datasets. In contrast, our model only requires identity annotation, and can achieve competitive performance. GLTR [31] employs dilated temporal convolutions to capture the multi-granular temporal dependencies, and achieves impressive results on these two datasets. However, GLTR does not consider spatial multi-granularity. This is why GLTR obtains less impressive results on MARS, where there tend to be misalignment in the sequence.

In summary, the above results have demonstrated the advantages of the proposed MGH model for video-based person re-ID. With only identity labels, MGH can achieve state-of-the-art performance on MARS, one of the largest existing public benchmarks for this task, in terms of mAP and top-1 accuracy. Meanwhile, MGH also achieves competitive results on iLIDS-VID and PRID-2011.

### 4.5. Results Visualization

We visualize some person re-ID results in Figure 6. As can be observed, it is difficult for the baseline model to distinguish people sharing similar appearances when there are misalignments and occlusions, resulting in relatively low top-1 accuracy. In these cases, the proposed MGH reduces the visual ambiguity by employing multi-granular spatial and temporal clues. At the same time, MGH achieves more robust results under different illumination conditions.

To better understand the attentive node feature aggregation module, the attention weights of two occlusion examples are visualized in Figure 7. On the one hand, for the global spatial partition, the images that contain a larger foreground person tend to be assigned with higher weights. On the other hand, for the local partitions, the parts belonging to the target person have obviously higher weights than the backgrounds. This indicates that the node attention module can adaptively concentrate on the discriminative parts, which validates the effectiveness of the attention module.

## 5. Conclusion

This paper proposed a multi-granular hypergraph learning framework to address video-based person re-ID. The proposed framework explicitly leveraged multi-granular spatial and temporal clues in the video sequence by learning a sophisticatedly-designed hypergraph neural network. In the learning process, we developed an attention mechanism to aggregate node-level features to yield more discriminative graph representations. In addition, we learned more diversified multi-granular features based on a novel mutual information loss. Extensive experiments were conducted on three person re-ID benchmarks, where the proposed framework achieved favorable performance compared with recent state-of-the-art methods.

## References


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 4

[2] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 2

[3] Song Bai, Peng Tang, Philip H. S. Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *CVPR*, 2019. 2

[4] Song Bai, Feihu Zhang, and Philip H. S. Torr. Hypergraph convolution and hypergraph attention. *CoRR*, abs/1901.08150, 2019. 3

[5] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 2018. 2, 7, 8

[6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep CRF for person re-identification. In *CVPR*, 2018. 2

[7] Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, and Ling Shao. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*, 2017. 2

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, 2019. 2

[9] Dahjung Chung, Khalid Tahboub, and Edward J. Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, 2017. 2

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[12] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*, 2015. 2

[13] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[14] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI*, 2019. 2, 3

[15] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. STA: spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, 2019. 7, 8

[16] Niloofar Gheissari, Thomas B. Sebastian, and Richard I. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006. 2

[17] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[19] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 5

[20] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 5

[21] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. VRSTC: occlusion-free video person re-identification. In *CVPR*, 2019. 6, 7

[22] Yuchi Huang, Qingshan Liu, and Dimitris N. Metaxas. Video object segmentation by hypergraph cut. In *CVPR*, 2009. 3

[23] Yuchi Huang, Qingshan Liu, Shaoting Zhang, and Dimitris N. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *CVPR*, 2010. 3

[24] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. In *ICML*, 2010. 2

[25] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *IJCAI*, 2019. 3

[26] Srikrishna Karanam, Yang Li, and Richard J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 2

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[28] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 2

[29] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew C. H. Lee, Ben Glocker, and Daniel Rueckert. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *MICCAI*, 2017. 2

[30] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 1

[31] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019. 2, 7, 8

[32] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019. 2

[33] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 2

[34] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018. 2, 7, 8

[35] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 2015. 2

[36] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 2

[37] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 2

[38] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 4

[39] Niall McLaughlin, Jesús Martínez del Rincón, and Paul C. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016. 2

[40] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017. 6

[41] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009. 2

[42] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, 2007. 2

[43] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018. 2

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[45] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, 2019. 7, 8

[46] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2

[47] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV*, 2018. 1

[48] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[49] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 1

[50] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 6

[51] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, 2018. 1, 3

[52] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014. 2, 5

[53] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 2

[54] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 6

[55] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 2

[56] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Qi Tian. Adaptive graph representation learning for video person re-identification. *CoRR*, abs/1909.02240, 2019. 7

[57] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018. 2, 7, 8

[58] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017. 7, 8

[59] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Pratim Talukdar. Hypergcn: Hypergraph convolutional networks for semi-supervised classification. *CoRR*, 2018. 3

[60] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016. 2

[61] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *CVPR*, 2016. 2

[62] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, 2018. 2

[63] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-Sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, 2019. 7, 8

[64] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, 2019. 1, 3

[65] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2, 5, 7

[66] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2

[67] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, 2017. 2

[68] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012. 2

[69] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *NeurIPS*, 2006. 3

[70] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017. 2, 7, 8