HSGNet: Object Re-identification with Hierarchical Similarity Graph Network

Fei Shen, Mengwan Wei, and Junchi Ren

Abstract—Object re-identification method is made up of backbone network, feature aggregation, and loss function. However, most backbone networks lack a special mechanism to handle rich scale variations and mine discriminative feature representations. In this paper, we firstly design a hierarchical similarity graph module (HSGM) to reduce the conflict of backbone and re-identification networks. The designed HSGM builds a rich hierarchical graph to mine the mapping relationships between global-local and local-local. Secondly, we divide the feature map along with the spatial and channel directions in each hierarchical graph. The HSGM applies the spatial features and channel features extracted from different locations as nodes, respectively, and utilizes the similarity scores between nodes to construct spatial and channel similarity graphs. During the learning process of HSGM, we utilize a learnable parameter to re-optimize the importance of each position, as well as evaluate the correlation between different nodes. Thirdly, we develop a novel hierarchical similarity graph network (HSGNet) by embedding the HSGM in the backbone network. Furthermore, HSGM can be easily embedded into backbone networks of any depth to improve object re-identification ability. Finally, extensive experiments on three large-scale object datasets demonstrate that the proposed HSGNet is superior to state-of-the-art object re-identification approaches.

Index Terms—Hierarchical similarity graph, object re-identification, deep learning.

I. INTRODUCTION

Object re-identification [1–11] focuses on retrieval objects of interest from massive traffic data, playing a great potential for intelligent transportation systems. With the development of classification tasks, the object re-identification community has become more attractive. Existing object re-identification networks usually utilize backbone networks in classification networks to extract object features. However, these backbone networks were initially designed for classification tasks and can not deal well with problems that are similar between intra-classes. Therefore, mining discriminative features are crucial in improving object re-identification performance.

Deep learning based object re-identification models [12–14] consist of backbone networks, feature aggregation architectures, and loss function. In backbone networks, lots of mature deep learning networks, such as ResNeSt [15], SeNet [16], and Res2Net [17] can be adopted in object re-identification. However, as shown in Fig. 1 (a), since object being captured by different cameras installed at different city locations, the images of different objects have very highly similar appearance features on the object’s re-identification task; On the contrary, in some local areas, such as the annual inspection sign of the vehicle’s front window, the lights, the hairstyle of pedestrians, and the shoe area have obvious differences. Therefore, how to mine the saliency of the object to emphasize the discriminative local area is of great significance for object re-identification.

In the re-identification research field, the straightforward approach [7, 8, 14, 18–22] is to divide vehicles or pedestrians into several regions/keypoints, and then study features independently on each regions/keypoints. Depending on whether additional annotations are required, we can subdivide the method into partition-based and partition-free methods. The partition-based methods denote partition with the help of additional information, e.g., multi-DNN fusion siamese neural network (MFSNN) [23] part-aligned representations (PAR) [18], local feature preservation (LFP) [7], partial attention and multi-attribute learning network (PAMLNet) [24] and SpindleNet [8]. The partition-free methods represent tutoring without additional information, e.g., contrastive attention module network (CAMNet) [25], part-based convolutional baseline (PCB) [20], efficient multi-resolution network (emrn) [9], hybrid pyramidal graph network (HPGN) [6], visibility-aware part model (VPM) [22]. However, both types of methods have obvious drawbacks for object re-identification. The partition-based methods [7, 8, 18, 19] face the problem of additional manual labeling and requiring a lot of computing resources. As shown in Fig. 1 (b), although partition-free methods [20–22] do not require extra manual annotation, their performance is vulnerable to pose and scale changes. Therefore, there are large intra-class variations in pedestrian and vehicle images, which cause object re-identification to be very challenging.

For that, this paper designs a hierarchical similarity graph module (HSGM) to reduce the conflict of backbone and classification networks. The designed HSGM first builds a rich hierarchical graph to mine the mapping relationships between global-local and local-local. The HSGM applies the spatial features and channel features extracted from different
locations as nodes, respectively, and utilizes the similarity scores between nodes to construct spatial and channel similarity graphs. During the learning process of HSGM, we utilize a learnable parameter to re-optimize the importance of each position, as well as evaluate the correlation between different nodes. In addition, we develop a novel hierarchical similarity graph network (HSGNet) by embedding the HSGM in the backbone network. Furthermore, HSGM can be easily embedded into backbone networks of any depth to improve object re-identification ability.

The HSGNet can effectively mine object features without additional human-assisted annotation. The main reasons are summarized as follows: (i) For highly similar appearance features of the different classes, the HSGNet can construct a node to represent a local region (e.g., lights or logos) in a similarity graph (i.e., spatial and channel). Then, it can learn a node’s feature representation from the node (e.g., lights or logos) and the node’s neighbors (e.g., roof and door). Since we consider both internal and external relationships between nodes, this enables each local node feature to be enhanced. In other words, we do not treat local features in isolation, but consider the mixed features of the local features and its surrounding nodes. (ii) For different pose and position changes of the same object, the HSGNet can dynamically optimize the most representative areas (i.e., local and global) under this background of the hierarchical graph, which focuses on the discriminative features. In other words, we need to pay more attention to the saliency or representative features of the object through this relationships of global-local and local-local.

The main contributions of this paper can be summarized as follows:

- A plug-and-play hierarchical similarity graph module (HSGM) is proposed to reduce the conflict of backbone and classification networks and aggregate the node’s features via spatial and channel directions.
- According to the proposed HSGM, we develop a novel hierarchical similarity graph network (HSGNet) to improve the performance for object re-identification. This is the first work that mines the important features via similarity graphs between the nodes to the best of our knowledge.
- We have conducted extensive experiments and achieved promising performance gains on three large-scale object datasets. Then, a large number of ablation studies also verify the effectiveness of the core mechanisms in the HSGNet for object re-identification.

II. RELATED WORK

Object re-identification progress is reviewed from two aspects: (1) feature extraction and (2) loss functions.

A. Feature Extraction

Due to the development of upstream backbone networks, such as ResNet [26], SeNet [16] and Res2Net [17], the re-identification community has grown very rapidly. Unfortunately, the re-identification task belongs to a kind of fine-grained image retrieval and inter-classes objects have highly similar appearance features. Moreover, the backbone networks cannot satisfactorily address the difference in local features since these methods are developed for classification tasks. For that, based on these classical backbone networks, [6, 7, 20] have developed many partition methods for object re-identification. Among them, we can subdivide the partition method into partition-based and partition-free methods. Ye et al. [27] used the attention mechanism on two-stream network to learn intra-modal difference and cross-modal relationship to improve feature alignment. The LFP [7] employs a fast R-CNN detector to narrow down the potential search area, and then uses the re-identification model to extract global features. Based on the two-stream network, Zhang et al. [28] separated the feature to three part and six part respectively, and attempted to align global and local features at the same time. The adaptive attention vehicle re-identification (AAVER) [29] employs key-point detection module to localize the local features and use adaptive key-point selection module to learning the relationship of parts.

The similar approach is also used in partition-free methods, which learn local features to achieve more detailed feature alignment. For example, the HPGN [6] proposes a pyramid graph network (PGN) to explore the spatial meaning of different regional features for vehicle re-identification. The PCB [20] utilizes a unified block strategy and uses independent convolution modules to extract features from each block for person re-identification. The quadruple directional deep learning networks (QD-DLF) [30] proposes four novelty directional pooling layers to resist the negative impact of viewpoint changes on vehicle re-identification by fusing features from different orientations.

B. Hierarchical Similarity Graph Module

C. Loss Functions

Loss functions for object re-identification can be divided into three types. (1) The identification loss function based on cross entropy loss [31], which aims to maximize the posterior probability of each object subject. The EMRN [9] proposes a multi-resolution features dimension uniform module and multi cross entropy losses to fix dimensional features from images of varying resolutions. (2) The classification loss function, such as contrastive loss [25, 32] function and hybrid similarity function [33], which is with the goal of setting similarities of positive pairs (i.e., object subjects of the same class label) are positive while making similarities of negative pairs (i.e., object subjects of different class labels) are negative. For example, the contrastive attention module network (CAMNet) proposes...
two different mode loss to assess one part feature’s importance based on all parts. (3) The ranking loss function [34, 35], which focuses on pulling distances of positive pairs as close as possible and pushing distances of negative pairs as far as possible. Recently object re-identification methods prefer to jointly use cross entropy loss [36] and triplet loss [34, 35] or its variants [37–47]. Since the cross entropy loss is beneficial to ensure a high convergence speed, while the triplet loss pulls distances of positive pairs as close as possible and pushes distances of negative pairs as far as possible.

### III. METHODOLOGY

As shown in Fig. 2, the hierarchical similarity graph module (HSGM) consists of hierarchically dividing, spatial & channel similarity graph learning, and hierarchically aggregation. The core idea is that we do not treat local features in isolation, but consider the mixed features of the local features and its surrounding nodes. The processing of the proposed HSGM is as follows.

1) Hierarchically Divide and Hierarchically Aggregation: Suppose there is a feature map, we first copy it three times and hierarchically divide it into $D_1$, $D_2$, and $D_3$, where $D_1$, $D_2$, and $D_3$ respectively have 1, 2, and 4 uniform feature maps. It is worth noting that these uniformly divided feature maps (i.e., $D_1$, $D_2$, and $D_3$) have different local features or global features. $D_1$ can represent the global features of the image. $D_2$ and $D_3$ can denote critical local features of the object. We construct $D_2$ and $D_1$, $D_2$, and $D_3$ into global-local and local-local pairs, respectively, to focus on the different importance features of the object. In other words, we need to pay more attention to the saliency or representative features of the object through this relationships of global-local and local-local. Then, seven feature maps are independently fed to the similarity graphs, which contain a spatial similarity graph (SSG) learning branch and a channel similarity graph (CSG) learning branch, to obtain more discriminative features. More detail about the SSG and CSG are described on Section III-2 and Section III-3. Furthermore, the hierarchical aggregation up-sample $D_3$ by two times and adds $D_2$, and the same on $D_2$ is up-sampled two times and adds $D_1$. Based on $D_1$, $D_2$, and $D_3$, the different feature maps are analogously fed to SSG and CSG independently, as shown in Figure 3.

2) Spatial Similarity Graph Learning: To simplify the description, assume that $X \in R^{h \times w \times c}$ is a three-dimensional feature map in one of the different feature maps, where $h$, $w$, and $c$ represent the height, width, and channel number of $X$. The specific operation steps of the spatial similarity graphs learning branch are as follows. We first perform channel squeezed on the feature map $X$, and the calculation process of channel squeezed is formulated as follows:

$$S_{i,j} = \frac{1}{c} \sum_{m} X_{i,j,m}, \quad (1)$$

where $X_{i,j,m}$ denotes represents the feature map value at row $i$, column $j$ on the $m$ channel, $S_{i,j}$ represents the feature map value of the corresponding spatial position after channel squeezed. Hence, we can obtain $S = \{S_{i,j} | i = 1, 2, 3, \ldots, h, j = 1, 2, 3, \ldots, w \} \in R^{h \times w}$. Then, according to Eq. (2), we apply three different average pooling (AP) strategies to feature map $S$.

$$S'_{k} = AP^k(S), \quad (2)$$

where $AP^k$ represents AP layer with $k \times k$ pixels, which set $k = 1, 2, 3$ in all experiments. Besides, we set $S' = \{S'_1, S'_2, S'_3\}$ to increase the perception domain at different areas.

According to the above $S'$, spatial similarity graphs $SSG_k = \{V_k, E_k\}$ are constructed to mine significant features. The vector $V_k$ and $E_k$ respectively denotes spatial node sets and spatial edge sets between different nodes. The SSG (i.e., $SSG_1$, $SSG_2$ and $SSG_3$) of single-scale are fused into $SSG_k$ of multi-scale through element-wise addition operation, as shown in Figure 3.

To better explain the principle, the subscript of $SSG_k$ and $X'_k$ are removed. We can reshape $X'_k$ as $X' \in R^{d'}$ with $d = h \times w$ features. Then, we contract each local feature as a node in HSGM. Therefore, we can regard an instance SSG of local
and the neighbor nodes of the node at the boundary are where channel node sets $V$ in Eq. (3). It is determined by the sliding window of the determination rule of channel edge sets $C$. According to Eq. (3), the SSG’s edges are built.

According to Eq. (7), we can obtain $C = \{C_m \mid m = 1, 2, 3, \ldots, e\} \in R^e$. Then, we construct a channel similarity graph (CSG), $CSG = \{V^c \in R^d, E^c \in R^{c \times c}\}$, where channel node sets $V^c \{C_m \mid m = 1, 2, 3, \ldots, e\}$ and the determination rule of channel edge sets $E^c$ is the same as Eq. (3). It is determined by the sliding window of $s$ strides, and the neighbor nodes of the node at the boundary are correspondingly reduced. The difference between $E^c$ and $E^s$ is that the former performs neighborhood determination in the channel dimension, while the latter conducts neighborhood determination in the spatial dimension. Then, based on the CSG, the features of each channel node are aggregated to obtain the $U^c = (I^c + E^c) V^c$ aggregated by the channel nodes, where $I^c$ is the identity matrix. Further, each channel aggregated node $A^c$ is weighted with a learnable weight parameter $\delta \in R^c$, as shown in Eq. (8), to mine the importance in the channel dimension.

$$A^c = U^c \odot \delta$$

The output $A^{ms}$ of the SSG and the output $A^c$ of the CSG are multiplied to obtain an enhanced feature map $A^{msc}$, as Eq. (9):

$$A^{msc} = A^{ms} A^c$$

Besides, to increase the SG’s convergence in the training phase, we enhance SG according to Eq. (10).

$$O = \text{LeakyReLU}(\text{BN}(A^{msc}))$$

where the BN and LeakyReLU respectively represent the batch normalization and the leaky rectified linear unit.

Lastly, inspired by residual mechanism [26], the $X^D$ is formulated as:

$$X^D = O + X.$$  

A. Combined with Backbone Network

As shown in Fig. 4, the hierarchical similarity graph module (HSGM) can be easily embedded into existing backbone network at any depth. Inspired by IANet [48], HSGM is placed at the down-sampling place of every residual group, which is the bottleneck of the block. Based on the proposed HSGM, a flexible hierarchical similarity graph network (HSGNet) is proposed for object re-identification. Specifically, For HSGNet based object re-identification, we adopt a ResNet50 as the backbone network. As shown in the Section IV for detailed analysis and comparison, our proposed HSGM is only inserted into the backbone network between stage3 and stage4 layers. Following the strong baseline model [49], the ’last stride=1’ training trick is used to set the last stride of the stage 4 down-sampling to 1 for retaining more spatial information in the learned feature maps. Besides, the HSGNet also contains a global max pooling (GPM) and a CBR composite unit. The CBR denotes a convolutional layer using $1 \times 1$ sized filters, a batch normalization layer [50] and a ReLU [51] activation function.

Therefore, as shown in Fig. 4, there are one triplet loss $L_{triplet}$ and one label smoothing cross entropy (LSCE) loss $L_{lsce}$ reconfigured to supervise the proposed HSGNet, thus the total loss function is formulated as follows:

$$L_{total} = \alpha L_{triplet} + \beta L_{lsce},$$

where $\alpha$ and $\beta$ are both a hype-parameter, used to control the balance between different losses. Their default value are
respectively set to 0.5 and 1.0 via cross-validation. The LSCE loss function $L_{\text{lsce}}$ is formulated as follows:

$$L_{\text{lsce}}(X, l) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \delta(l_i, j) \log \left( \frac{e^{W_j^T X_i}}{\sum_{k=1}^{K} e^{W_k^T X_i}} \right),$$

(13)

where $X$ and $l$ are respectively a training set and the class label information; $N$ is the number of training samples; $K$ is the number of classes; $W = [W_1, W_2, W_3, ..., W_K]$ is a learnable parameter matrix. The $(X_i, l_i)$ is the $i$-th training sample and $l_i \in \{1, 2, 3, ..., K\}$.

The triplet loss function $L_{\text{triplet}}$ is formulated as follows:

$$L_{\text{triplet}}(X^a, X^n, X^p) = -\frac{1}{N} \sum_{i=1}^{N} \max(\|X_i^a - X_i^n\|_2 - \|X_i^a - X_i^p\|_2 - \delta, 0),$$

(15)

where $M$ and $(X^a, X^n, X^p)$ are respectively a set of training triplets and the number of training triplets. The $\|\cdot\|_2$ denotes the Euclidean distance. For the $i$-th training triplet, $(X_i^a, X_i^n)$ and $(X_i^a, X_i^p)$ are a positive pair of the same class label and a negative pair of different class labels, respectively. The margin constant of triplet loss $L_{\text{triplet}}$ (i.e., $\delta$ of Eq. 15) and the smoothing regularization degree of label smoothing cross entropy loss $L_{\text{lsce}}$ (i.e., $\gamma$ of Eq. 14) are respectively set to 1.2 and 0.1 in all experiments.

IV. EXPERIMENTS AND ANALYSIS

In this section, we evaluate our HSGNet on three well-known object re-identification datasets, namely, Market-1501 [52], VeRi776 [53], and VehicleID [38]. For performance metrics, commonly-used mean average precision (mAP) [38, 52, 53], and cumulative match characteristic (CMC) [38, 52, 53] curve are applied.

A. Datasets

**Market-1501** [52] is a pedestrian Re-ID dataset captured by six cameras. It contains 32,668 pedestrian images of 1,501 identities. The training subset includes 12,936 images of 751 identities, while the test subset holds images of the rest 750 identities, i.e., 19,732 gallery pedestrian images and 3,368 query pedestrian images.

**VeRi776** [53] is a vehicle Re-ID dataset captured from real traffic scenarios by using 20 surveillance cameras. It totally consists of 776 subjects. The training subset contains 37,746 images of 576 subjects, and the rest 200 subjects are applied for the testing subset. Furthermore, the test subset includes a probe subset of 1,678 images and a gallery subset of 11,579 images.

**VehicleID** [38] is constructed from video sequences captured by cameras in different locations, backgrounds and lighting conditions, such as highway intersections, road crossings, parking lots, etc, which totally includes 221,763 images of 26,267 classes. It has three test subsets, each of which randomly selects one image of a tested object to form the probe subset, and all remaining images form the gallery subset.

B. Implementation Detail

The operating system used in this experiment is Ubuntu16.04 and the deep learning toolbox is Pytorch with two V100 GPUs on the same machine for fair comparison. Training configurations are summarized as follows. (1) Common object re-identification training tricks are used, such as using ImageNet pretrained weights, random erasing, random flipping, and z-score normalization. (2) We respectively resize the image resolution uniformly to 256 x 256 and 256 x 128 on vehicle (i.e., VeRi776 and VehicleID) and person (i.e., Market-1501) datasets. (3) The mini-batch stochastic gradient descent (SGD) method is applied to train parameters. The weight decays are set to be $5 \times 10^{-4}$, and the momentums are set to be 0.9. There are 120 epochs for the training process. The learning rates are initialized to 0.001, and linearly warmed up to 0.01 in the first 10 epochs. After warming up, the learning rates are maintained at 0.01 from 11st to 30th epochs. Then, the learning rates decays by 10% every 20 epochs. (4) Each mini-batch includes 32 subjects and each subject holds 4 images. In the testing phase, the 2048-dimensional feature output by the GMP layer in Fig. 4 is used as the final image feature, and the cosine distance is used as the similarity measure for object re-identification.

C. Comparison with State-of-the-art Methods

Comparisons on Market-1501. As shown in Table I, the proposed hierarchical similarity graph network (HSGNet) obtains the highest 95.0% Rank1, 98.9% Rank5 and 86.6% mAP. For example, compared with two partition-based methods, i.e., SpindleNet [8] and PAR [18], the HSGNet is higher than 17.9% and 13.8% on Rank1, respectively. Considering the partition-free methods, i.e., AANet [58] and IANet [48], it is still defeated 4.2% and 3.5% on mAP, respectively. Although our method is slightly lower on Rank1 than P²-Net [61], which uses an additional semantic segmentation model, we defeat it by 1.0% on mAP. Therefore, integrating Rank and mAP retrieval comparison results into account, our method remains competitive on Market-1501.

| Methods          | Rank1  | Rank5  | mAP     |
|------------------|--------|--------|---------|
| SpindleNet [8]   | 76.9   | 91.5   | 63.4    |
| PAR [18]         | 81.0   | 92.0   | 66.1    |
| LSRO [54]        | 84.0   | -      | 73.1    |
| MultiScale [55]  | 91.2   | 75.7   | -       |
| HA-CNN [56]      | 91.4   | 76.6   | -       |
| DuATM [59]       | 91.2   | 76.8   | -       |
| PCB [20]         | 93.0   | 80.8   | -       |
| AANet [58]       | 93.8   | 82.4   | -       |
| HPM [21]         | 94.2   | 97.5   | 82.7    |
| HORerID[59]      | 94.2   | 84.9   | -       |
| SNR [60]         | 94.4   | 84.7   | -       |
| IANet [48]       | 94.4   | -      | 83.1    |
| P²-Net [61]      | 95.2   | 98.2   | 85.6    |
| BoT (Baseline)   | 93.8   | 97.3   | 84.5    |
| HSGNet           | 95.0   | 98.9   | 86.6    |
Comparisons on VeRi776. The performance of HSGNet and multiple state-of-art methods are compared shown in Table II. Firstly, Compared with those SOTA methods, the proposed HSGNet obtains the best Rank1 (i.e., 96.94%) and the highest mAP (i.e., 80.65%). Secondly, Compared with multi-scale methods, the HSGNet defeats the EMRN [9] by 2.19% in terms of Rank1 and 0.87% in terms of mAP. Thirdly, compared with partition-based methods (i.e., MRM [64] and part regularization [65]), the proposed HSGNet still defeat 5.17% and 2.64% on Rank1, respectively. Lastly, the HSGNet still respectively outperforms the method of additional vehicle attribute annotations, i.e., Appearance+License [66] by 2.57% and 1.53% larger on mAP and Rank1 identification rate. It is worth noting that the proposed HSGNet method does not require license plates. Therefore, our proposed HSGNet is superior than these state-of-the-art approaches on VeRi776 dataset in both partition-based and partition-free methods.

Comparison on VehicleID In fact, the VehicleID [38] dataset has a larger data scale than the VeRi776 [53] dataset. However, the proposed HSGNet method still can obtain the 1st place and outperforms those state-of-the-art methods under comparison, as occurred on the VeRi776 dataset, as shown in Table II. For example, the proposed HSGNet method is the only case that acquires a more than 82% rank-1 identification rate on the Test800 subset, which is respectively 3.22% higher than that of the 2nd places (i.e., Appearance+License [66]). Compared with the RAM [31] of the best partition-free method, the proposed HSGNet respectively defeats it by 7.57%, 5.51%, and 8.95% in terms of Rank1 identification rate on Test800, Test1600, and Test2400 subset. Moreover, On the largest Test 2400, compared with partition-based methods (i.e., MRM [64] and Part Regularization [65]), the proposed HSGNet still defeat 5.79% and 2.45% on Rank1 identification rate, respectively.

From Table I and Table II, the proposed HSGNet method obtains state-of-the-art performance on Market-1501, VeRi776, and VehicleID, which shows the effectiveness and generalization of our method. These results demonstrate that HSGNet can not only handle various scale and pose changes but also mine discriminative features.
D. Ablation Studies and Analysis

In previous subsection, we have shown the superiority of HSGNet via comparing to state-of-the-art methods. In what’s follow, we comprehensively analyze intrinsic factors that lead to HSGNet’s superiority, including: (1) The role of hierarchical similarity graph module (HSGM). (2) The effectiveness on which stage to plug the HSGM. (3) The effect of sampling strategy of the node in HSGM. (4) The universality for different backbones. (5) The comparisons of parameters and computations.

The role of hierarchical similarity graph module (HSGM). We conduct ablation experiments to analyze the effectiveness of different components of the hierarchical similarity graph network (HSGNet) on VeRi776 and Market-1501 datasets. We gradually increase each component to HSGM, respectively, including the HSGM w/o Spatial and Channel Similarity Graph, HSGM w/o Residual (Res.) Connection, HSGM w/o Spatial Similarity Graph, HSGM w/o Channel Similarity Graph, and HSGNet, respectively. Second, HSGM w/o Spatial Similarity Graph and HSGM w/o Channel Similarity Graph are also better than HSGM w/o Spatial and Channel Similarity Graph in all evaluation indicators, indicating that Spatial and Channel Similarity Graph can improve and enhance the feature map for object re-identification. Third, HSGNet also outperforms HSGM w/o Hierarchical and HSGM w/o Spatial and Channel Similarity Graph, which suggests that HSGNet can handle the difficulties of variable scales and positions.

Besides, we visualize the class activation maps of some examples from VeRi776 datasets, as shown in Figure 5. The purpose of HSGM is to mine more discriminative and recognizable features to improve the performance of object re-identification. The grad class activation maps (Grad-CAM) [67] visualized show that HSGNet can help the model pay more attention to local details, especially the lights, inspection signs, accessories, these are indeed differentiated information.

The effectiveness on which stage to plug the HSGM. We develop four methods that embed proposed HSGM at different stages of the ResNet-50’s bottleneck layer. The S1, S2, S3, and S4 respectively embed HSGM behind stage1, stage2, stage3, and stage4. The corresponding results are shown in Fig. 6 and Fig. 7. It is observed that no matter where the HSGM is embedded in the ResNet-50, there is a performance benefit. For example, according to Fig. 6 (a), the proposed HSGM respectively brings 0.8%, 0.9%, 1.2%, and 1.0% performance improvement than Baseline method from stage1 (S1) to stage4 (S4) on Rank1 re-identification rate of the Market-1501 dataset. As shown in Fig. 7 (b), the proposed HSGM respectively brings 2.4%, 2.8%, 3.3%, and 3.0% performance improvement than Baseline method from stage1 (S1) to stage4 (S4) on mAP performance of the VeRi776 dataset. Especially, whether it is Rank1 re-identification or mAP performance, the performance improvements of HSGM in stage3 (S3) are more extensive on two datasets simultaneously.

The effect of sampling strategy of the node in HSGM. The HSGM is a key component of HSGNet method, thus the influence of sampling strategy of Node in HSGM is evaluated. Recalling our proposed method, the default HSGNet adopts $k_1=1$, $k_2=2$, and $k_3=3$ sampling strategy of node. Likewise, as shown in Table IV, we design HSGMs with three different configurations, such as (1) $k_1=1$, $k_2=1$, and $k_3=1$, (2) $k_1=2$, $k_2=2$, and $k_3=2$, (3) $k_1=3$, $k_2=3$, and $k_3=3$, namely, HSGNet1, HSGNet2, and HSGNet3, respectively.

From Table IV, the HSGNet method of multi-scale sampling sizes consistently outperforms HSGNet1, HSGNet2, and HSGNet3 methods those of fixed sampling sizes on three two datasets. For example, on the VeRi776 dataset, HSGNet’s mAP is 1.05%, 0.84% and 0.67% higher than that of HSGNet1, HSGNet2, and HSGNet3 methods, respectively. On the Market-1501 dataset, Git method defeats HSGNet1, HSGNet2, and...
This paper proposes a novel hierarchical similarity graph network (HSGNet) with a hierarchical similarity graph module (HSGM) for object re-identification. Specifically, HSGM constructs a hierarchical graph to explore the pairwise relationships between global-local and local-local. Then, in each hierarchical graph, the HSGM respectively regards spatial features and channel features extracted from different locations as nodes and utilizes the similarity scores between nodes to construct a spatial similarity graph and a channel similarity graph. Besides, we construct extensive experiments to analyze and discuss our proposed HSGNet, including: (1) comparing a lot of state-of-the-art approaches to demonstrate the HSGNet method’s superiority; (2) the ablation study of the HSGM; (3) the universality of HSGM to different backbones. (4) the comparison of parameters and computations. Experiments on three large-scale datasets demonstrate that the proposed HSGNet outperforms state-of-the-art object re-identification approaches.

### REFERENCES

[1] X. Zhang and B. Bhanu, “Unbiased spatio-temporal representation with uncertainty control for person re-identification,” IEEE Transactions on Cognitive and Developmental Systems, 2022.

[2] Y. Huang, Q. Wang, W. Yang, Q. Liao, and H. Meng, “Hi-Part: Occluded person re-identification with part-discovery: Occluded person re-identification with part-detection,” IEEE Internet of Things Journal, 2017.

[3] W. Zhu, R. Hu, Z. Wang, D. Li, and X. Gao, “Tell the truth from the front: Anti-disguise vehicle re-identification,” in 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020, pp. 1–6.

[4] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, “Spindle net: Person re-identification with human body region guided feature decomposition and fusion,” pp. 907–915, 2021.

[5] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, “Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification,” IEEE Transactions on Intelligent Transportation Systems, 2021.

[6] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, “Modeling context appearance changes for person re-identification via ipscgn,” Neurocomputing, vol. 483, pp. 19–27, 2021.

[7] C. Liu, H. Yang, J. Zhu, Q. Zhou, and S. Zheng, “Message passing with graph convolutional network for person re-identification,” Neurocomputing, vol. 423, pp. 57–70, 2021.

[8] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, “An efficient multi-resolution network for vehicle re-identification,” IEEE Internet of Things Journal, 2021.

[9] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, “Diverse part discovery: Occluded person re-identification with part-
aware transformer,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2898–2907.

[11] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, “Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8450–8459.

[12] A. M. N. Taufique and A. Savakis, “Labnet: Local graph aggregation network with class balanced loss for vehicle re-identification,” Neurocomputing, vol. 463, pp. 122–132, 2021.

[13] F. Shen, X. Peng, L. Wang, X. Zhang, M. Shu, and Y. Wang, “Hsgm: A hierarchical similarity graph module for object re-identification,” in 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.

[14] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, “Gt: Graph interactive transformer for vehicle re-identification,” arXiv preprint arXiv:2107.05475, 2021.

[15] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Mamatha et al., “Resnet: Split-attention networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.

[16] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[17] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2net: A new multi-scale backbone architecture,” IEEE TPAMI, 2020.

[18] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3219–3228.

[19] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, “Pantri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data,” in International Conference on Computer Vision, 2019, pp. 211–220.

[20] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in European Conference on Computer Vision, 2018, pp. 480–496.

[21] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, “Horizontal pyramid matching for person re-identification,” in AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8295–8302.

[22] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, “Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 393–402.

[23] C. Cui, N. Sang, C. Gao, and L. Zou, “Vehicle re-identification by fusing multiple deep neural networks,” in IEEE International Conference on Image Processing Theory, Tools and Applications, 2017, pp. 1–6.

[24] S. Tumrani, Z. Deng, H. Lin, and J. Shao, “Partial attention and multi-attribute learning for vehicle re-identification,” Pattern Recognition Letters, vol. 138, pp. 290–297, 2020.

[25] M. Li, M. Wei, X. He, and F. Shen, “Enhancing pary features via contrastive attention module for vehicle re-identification,” in Conference on International Conference on Image Processing. IEEE, 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[27] M. Ye, J. Shen, D. J Crandall, L. Shao, and J. Luo, “Dynamic dual-attentive aggregation learning for visible-infrared person re-identification,” in European Conference on Computer Vision. Springer, 2020, pp. 229–247.

[28] L. Zhang, G. Du, F. Liu, H. Tu, and X. Shu, “Global-local multiple granularity learning for cross-modality visible-infrared person reidentification,” IEEE Transactions on Neural Networks and Learning Systems, 2021.

[29] P. Khorrambashahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, “A dual-path model with adaptive attention for vehicle re-identification,” in International Conference on Computer Vision, 2019, pp. 6132–6141.

[30] J. Zhu, H. Zeng, J. Huang, S. Liao, L. Zhen, C. Cai, and L. Zheng, “Vehicle re-identification using quadruple directional deep learning features,” Transactions on Intelligent Transportation Systems, vol. 21, no. 1, pp. 410–420, 2020.

[31] X. Liu, S. Zhang, Q. Huang, and W. Gao, “Ram: a region-aware deep model for vehicle re-identification,” in International Conference on Multimedia & Expo, 2018, pp. 1–6.

[32] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in Conference on Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.

[33] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, “Deep hybrid similarity learning for person re-identification,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3183–3193, 2018.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

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

[36] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.

[37] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram, “Vehicle re-identification: an efficient baseline using triplet embedding,” in International Joint Conference on Neural Networks, 2019, pp. 1–9.

[38] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, “Deep relative distance learning: Tell the difference between similar vehicles,” in Conference on Computer Vision and Pattern Recognition, 2016, pp. 2167–2175.

[39] X. Liu, S. Zhang, Y. Wang, and Q. Tian, “Group-group loss based global-regional feature learning for vehicle re-identification,” IEEE Transactions on Image Processing (Early Access), 2019.

[40] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, “Group-sensitive triplet embedding for vehicle reidentification,” IEEE Transactions on Multimedia, vol. 20, no. 9, pp. 2385–2399, 2018.

[41] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, “Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles,” in International Conference on Computer Vision, 2017, pp. 562–570.

[42] J. Hou, H. Zeng, J. Zhu, J. Hou, J. Chen, and K.-K. Ma, “Deep quadruplet appearance learning for vehicle re-identification,” IEEE Transactions on Vehicular Technology, vol. 68, no. 9, pp. 8512–8522, 2019.

[43] H. Guo, K. Zhu, M. Tang, and J. Wang, “Two-level attention network with multi-grain ranking loss for vehicle re-identification,” IEEE Transactions on Image Processing, pp. 4328–4338, 2019.

[44] H. Guo, C. Zhao, Z. Lui, J. Wang, and H. Lu, “Learning coarse-to-fine structured feature embedding for vehicle re-identification,” in AAAI Conference on Artificial Intelligence, 2018, pp. 6853–6860.

[45] Y. Zhou, L. Shao, and A. Dhab, “Viewpoint-aware attentive multi-view inference for vehicle re-identification,” in Conference on Computer Vision and Pattern Recognition, 2018, pp. 6489–6498.
[46] W. Lin, Y. Li, X. Yang, P. Peng, and J. Xing, “Multi-view learning for vehicle re-identification,” in International Conference on Multimedia & Expo, 2019, pp. 832–837.

[47] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, “Vehicle re-identification with viewpoint-aware metric learning,” in International Conference on Computer Vision, 2019, pp. 8282–8291.

[48] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “Interaction-and-aggregation network for person re-identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9317–9326.

[49] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, “A strong baseline and batch normalization neck for deep person re-identification,” IEEE Transactions on Multimedia (Early Access), 2019.

[50] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning, 2015, pp. 448–456.

[51] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in International Conference on Machine Learning, 2010, pp. 807–814.

[52] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1116–1124.

[53] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in European Conference on Computer Vision, 2016, pp. 869–884.

[54] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762.

[55] Y. Chen, X. Zhu, and S. Gong, “Person re-identification by deep learning multi-scale representations,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2590–2600.

[56] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2285–2294.

[57] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 542–551.

[58] C.-P. Tay, S. Roy, and K.-H. Yap, “Aanet: Attribute attention network for person re-identifications,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7134–7143.

[59] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, J. Sun et al., “High-order information matters: Learning relation and topology for occluded person re-identification,” in Conference on Computer Vision and Pattern Recognition, 2020.

[60] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, “Style normalization and restitution for generalizable person re-identification,” in Conference on Computer Vision and Pattern Recognition, 2020.

[61] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, “Beyond human parts: Dual part-aligned representations for person re-identification,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3642–3651.

[62] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269.

[63] X. Yang, C. Lang, P. Peng, and J. Xing, “Vehicle re-identification by multi-grain learning,” in International Conference on Image Processing, 2019, pp. 3113–3117.

[64] J. Peng, H. Wang, T. Zhao, and X. Fu, “Learning multi-region features for vehicle re-identification with context-based ranking method,” Neurocomputing, vol. 359, pp. 427–437, 2019.

[65] B. He, J. Li, Y. Zhao, and Y. Tian, “Part-regularized near-duplicate vehicle re-identification,” in Conference on Computer Vision and Pattern Recognition, 2019, pp. 3997–4005.

[66] Y. He, C. Dong, and Y. Wei, “Combination of appearance and license plate features for vehicle re-identification,” in International Conference on Image Processing, 2019, pp. 3108–3112.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in International Conference on Computer Vision, 2017, pp. 618–626.