Progressive Context-aware Graph Feature Learning for Target Re-identification

Min Cao, Cong Ding, Chen Chen*, Hao Dou, Xiyuan Hu and Junchi Yan, *Member, IEEE*

***Abstract*—This paper aims at robust and discriminative feature learning for target re-identification (Re-ID). In addition to paying attention to the individual appearance information as in most Re-ID methods, we further utilize the abundant contextual information as additional clues to guide the feature learning. Graph as a format of structured data is used to represent the target sample with its context. It describes the first-order appearance information of the samples and the second-order topological relationship information among samples, based on which we compute the feature representation by learning a graph feature embedding. We provide a detailed analysis of graph convolutional network mechanism applied in target Re-ID and propose a novel progressive context-aware graph feature learning method, in which the message passing is dominated by a pre-defined adjacency relationship followed by a learned relationship in a self-adaptive way. The proposed method fully exploits and utilizes contextual information at a low cost for Re-ID. Extensive experiments on five Re-ID benchmarks demonstrate the state-of-the-art performance of the proposed method.**

***Index Terms*—Target re-identification, graph convolutional network, feature learning, contextual information, graph feature learning.**

I. Introduction

Target Re-ID is a task of retrieving the same identity images across different camera views [1], [2]. Common target Re-ID task typically includes person Re-ID [3] and vehicle Re-ID [4], both of which have extensive application value in the intelligent surveillance system. Despite the promising progress made in recent years, target Re-ID remains a challenging task, mainly in the complex within-class variations caused by different views and the subtle between-class discrepancy caused by similar appearance.

Taking person Re-ID as an example, there are usually two basic steps to solve this task: feature learning [5], [6], [7], [8] and metric learning [5], [9], [10],[11]. A variety of methods are proposed to focus on some of these steps to improve Re-ID performance [12], [8], [13],[14]. The mainline of works concentrates on exploring individual appearance features [8], [15], [16] for Re-ID, as shown in Fig. 1 (a). These works seek to obtain the salient feature representation of the image and address misalignment of the image across views by the traditional algorithm [17], [12] or the deep learning technology [18], [19]. However, due to large within-class variance and small between-class variance in Re-ID, the individual appearance information is not powerful enough to distinguish different identities for Re-ID.

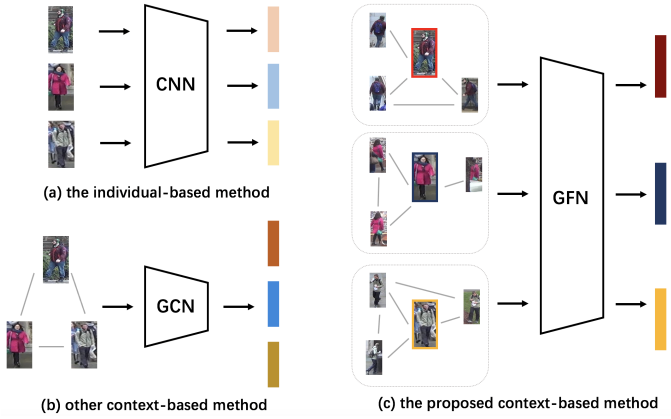

Fig. 1. Illustration of different kinds of Re-ID methods. The filled rectangle represents the sample's feature representation. (a): The features of three images are extracted independently based on the individual appearance information by the convolutional neural network (CNN); (b): The features of three images are extracted together with the consideration of the appearance information of other samples by the graph convolutional network (GCN) and the node feature is used as the image feature; (c): The features of three images are extracted independently with the consideration of the appearance information of other samples and the topological relationship information among samples by the proposed graph feature network (GFN) and the graph feature is used as the image feature.

In this context, another line of analysis is towards exploring richer contextual information as additional valuable clues to achieve Re-ID. The spatiotemporal information as a kind of contextual information is leveraged to complement the individual appearance information in target Re-ID [20], [21]. However, the video data needs to be processed in advance to obtain the spatiotemporal information. Alternatively, the contextual information associated with other samples in the dataset can be exploited relatively cheaply and easily by setting specific assumptions. The re-ranking Re-ID methods [22], [12] are proposed to refine the ranking list by utilizing such

Chen Chen is the corresponding author (email: chen.chen@ia.ac.cn).
Min Cao and Cong Ding are with School of Computer Science and Technology, Soochow University, Suzhou 215006, China (email: mcao@suda.edu.cn).
Chen Chen and Hao Dou are with the Institute of Automation Chinese Academy of Sciences, Beijing 100190, China.
Xiyuan Hu is with School of Computer Science and Engineering, Nanjing University of Science and Technology.
Junchi Yan is with Department of Computer Science and Engineering, and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, China.
This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes the further discussions about the proposed method. This material is 207 KB in size."
This work is supported by the National Science Foundation of China under Grant NSFC 62002252, and is also partially supported by the National Science Foundation of China under Grant NSFC 61906194, Collaborative Innovation Center of Novel Software Technology and Industrialization, and Liaoning Collaboration Innovation Center For CSLE.

contextual information in an unsupervised way. For example, the method in [22] refines the ranking results by encoding and comparing k-reciprocal nearest neighbors between the samples. Without training under the supervision of the labeled samples, these methods do not take full advantage of contextual information. Correspondingly, some researchers propose to learn the sample's feature representation by the aid of its contextual information from other samples in the dataset in a supervised manner [2], [23], [24], [25].

Graph as a kind of non-Euclidean structured data encodes a set of samples (*i.e.*, nodes) and their relationships (*i.e.*, edges) and naturally conveys the contextual information. The graph convolutional networks (GCNs) are correspondingly designed to solve the problem on the graph data and can also be adopted to learn the context-aware feature representation for Re-ID [2], [23], [20]. These methods focus on the graph's construction by proposing the complicated custom-built strategies from different perspectives. The node's feature representation is computed by GCNs and is used as the sample's representation, as shown in Fig. 1 (b). Although various attempts along this direction have led to great performance advances, there are still some shortcomings and limitations in these methods. Firstly, the sample is usually modeled as the node in the graph and, its context-aware feature representation is obtained by learning the node feature. They do not take full advantage of the high-order topological information in the graph for Re-ID. Secondly, these methods simply apply the GCN technology to learn the node feature for Re-ID, and the essential mechanism of GCNs is not discussed in depth. In fact, it is proved that the message passing among nodes is a critical component for the performance benefit in GCNs [26], [27], which however is not paid sufficient attention on these GCNs-based Re-ID methods [2], [23], [20]. Specifically, the sample's context and the adjacency relationships between samples dominate the message passing process. However in these methods, the context is usually obtained based on a subset of samples with relatively weak semantic associations and in a local contextual level (*i.e.*, a mini-batch under random shuffle), and the adjacency relationships between samples are obtained by measuring their similarities only based on a single dimension. All of these result in insufficient exploitation of the context and the relationship, which has a direct impact on the message passing and leads to a suboptimal solution. In addition, the over-smoothing due to multiple message passing processes is a common problem in GCNs and may affect the expressive power of GCNs to some extent [27], nevertheless, most of GCNs-based Re-ID methods ignore the problem.

In this paper, we aim to learn the target sample's context-aware feature representation with a high-level exploration and utilization of contextual information and low model complexity. To overcome the above-mentioned problems, we capture the sample's global context by searching its nearest neighbors in the whole dataset, and represent the sample with its global context by a graph data and compute the graph feature representation as the sample representation, as shown in Fig. 1 (c). At this point, the target Re-ID problem is transformed into a graph matching problem. In this manner, we strengthen the utilizations of the first-order appearance information of samples and the second-order topological relationship among them. For graph feature learning, considering that the majority of the benefit derives from the message passing process in GCNs, we revise the classical GCN framework by highlighting the message passing process and simplifying the non-linear mapping process. For this, we propose a progressive context-aware graph feature learning method by deriving a progressive message passing process and removing the non-linear mapping processes between layers in GCNs. By doing so, the over-smoothing problem is also mitigated.

The contributions of this paper are summarized as follows:

1) To obtain the discriminative feature representation, we abstract the sample with its global context as a structured data in the form of graph, and the sample's feature representation is extracted by learning the graph representation. We transform the target Re-ID problem into a graph matching problem.

2) For graph feature learning, we explore the GCN mechanism applied in target Re-ID in detail, and propose a progressive context-aware graph feature learning method to mine and utilize the contextual information thoroughly.

3) The proposed method can be readily applied to most existing offline Re-ID baselines to boost performance with its straightforward and simple framework. The extensive experiments on five Re-ID datasets verify the effectiveness and efficiency of the proposed method.

II. Related work

We propose a novel graph feature learning method to extract the sample's context-aware feature representation for Re-ID. In this section, we introduce the related works on the feature learning-based Re-ID methods and the GCN-based Re-ID methods, respectively.

A. Feature Learning for Target Re-ID

Feature learning plays a significant role in the targe Re-ID task. The existing feature learning-based methods mainly aim at extracting three types of features: global features extracted from the whole target image [28], [16], [29], local features generated from the local image regions [30], [15], [7], [31], [32],[33], and a fusion of global and local features [8], [34], [35], [36], [37],[38]. For example, Li *et al.* [16] focused on the joint learning of soft pixel attention and hard regional attention to obtaining the sample's global feature for person Re-ID; Guo *et al.* [6] applied a human parsing model and a self-attention mechanism to exploit both the accurate human parts and the coarse non-human parts for extracting the local features in person Re-ID. Meng *et al.* [39] proposed to align the view-aware feature representations by learning a fine-grained feature representation for vehicle Re-ID. These methods fully exploit the individual appearance information for feature learning and the abundant contextual information at different semantic levels is ignored, hindering the performance enhancement.

To overcome the limitation and further improve performance, researchers propose to explore the contextual information for extracting better feature representation [34], [40], [41], [24], [42], [43], [23]. In these methods, the contextual information is explored from three different semantic levels:

the information of the physical neighbors in the time-space dimension [44], [20], [21], the information of the k-nearest neighbors in the feature space [12], [24], [43], [23], [45] and the appearance information of other local image regions [46], [47]. For example, Yang *et al.* [21] proposed a novel spatial-temporal graph convolutional network to model the temporal relations of different frames and the spatial relations within a frame for video-based person Re-ID; Liu *et al.* [45] utilized the sample's contextual information from its high-order k nearest neighbors for computing the similarity between samples; Liu *et al.* [46] explored the vehicle parsing to learn discriminative part-level features and model the correlation among vehicle parts via graph convolutional networks, thus achieving precise part alignment for vehicle Re-ID. In this paper, we aim at mining the contextual information on the second semantic level, *i.e.,* exploring the information of the k-nearest neighbors in the feature space.

B. GCNs and its Application in Target Re-ID

Graph as a kind of non-Euclidean structured data models a set of objects and their relationships, and many real-world irregular data can be represented by the graph. Inspired by convolutional neural networks (CNNs) for processing regular structured data, GCNs [48], [49] are emerged to operate on the graph data with the node-focused application [50], [51] and the graph-focused application [52], [53].

Recent studies are proposed to solve Re-ID task by means of GCNs. For person Re-ID methods, according to the application scenario, we divide them into three categories: video-based person Re-ID, occluded person Re-ID and image-based person Re-ID. In video-based person Re-ID, the dependencies of time-space dimension among samples are exploited by constructing the graph and applying GCNs on the graph [34], [21], [40]. In occluded person Re-ID, researchers aim to align a set of local features across images by viewing it as a graph matching problem with the application of GCNs [41]. In image-based person Re-ID, the graph is constructed with nodes modeling image samples and edges modeling relationships among them, and the node features are learned by GCNs [24], [20], [42], [43], [23], [54], [2]. For vehicle Re-ID methods, researchers usually model the vehicle's appearance structure as a graph, its feature representation is learned by GCNs [46], [47].

As most related works with the proposed method, the image-based Re-ID methods focus on constructing graph data with a simple application of GCNs. These methods developed the complicated custom-built strategies for constructing the graph, with inadequate exploration of contextual information and the over-smoothing problem. They result in a suboptimal solution with high model complexity. By contrast, we focus on the analysis of the GCN mechanism in target Re-ID and propose a novel simple but powerful progressive context-aware graph learning method for Re-ID, in which the contextual information is fully exploited at a low cost and the over-smoothing problem is also effectively mitigated. We present a more detailed comparison with these related methods in Section III-C.

III. Methodology

Given a training set with labeled samples, we aim to learn a discriminative feature embedding function under the guidance of these samples and their context samples. In the testing phase, given a probe sample and a test set of gallery samples, we extract the context-aware feature representations of samples based on the learned feature embedding function. Target Re-ID task is achieved by computing the distance values between feature representations and then ranking the gallery samples based on these values.

We express the sample as a graph data, which is constructed by the sample and its context. The graph contains the appearance information of the sample and its context (*i.e.*, nodes) and also encodes the relationship information between them (*i.e.*, edges). At this point, we compute the sample's feature representation by learning the graph's feature embedding based on the GCN technology.

In the following sections, we first give a brief preliminary introduction on GCNs in Section III-A. We then elaborate on the proposed method in Section III-B. In Section III-C, we take a step further by discussing the differences with other related methods in detail.

A. Preliminary Introduction on GCNs

We denote a graph as $\mathcal{G} = (V, E)$ with the node feature matrix $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$, where $V = \{v_1, \cdots, v_n\}$ is a set of nodes, $E = \{e_{ij} | v_i, v_j \in V\}$ is a set of edges, the node feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ collects the nodes' initial feature representations and the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ digitizes the relations between nodes.

The GCNs-based methods usually involve two kinds of applications: node-focused application and graph-focused application.

In the node-focused application, it concentrates on computing a new node feature matrix. There is thus only a node embedding stage in which each node of a graph iteratively aggregates the feature representations of itself and its connected nodes (*i.e.*, its neighbors) in the graph to compute its new feature representation. This stage can be formatted as:

$$\mathbf{x}_i^{(r+1)} = \Psi_{\mathbf{\Theta}} \left(\mathbf{x}_i^{(r)}, \Gamma_{\bar{\mathbf{\Theta}}} (\mathbf{x}_j^{(r)} | e_{ij} \in E) \right), \quad r = 1, \cdots, R, \tag{1}$$

where $\mathbf{x}_i^{(r)}$ and $\mathbf{x}_j^{(r)}$ denote the feature vectors of nodes v_i and v_j at the r-th layer with $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(1)}$ being the initial feature vectors in $\mathbf{X}$, $\Psi_{\mathbf{\Theta}}$ and $\Gamma_{\bar{\mathbf{\Theta}}}$ are the non-linear functions with the learnable parameters $\mathbf{\Theta}$ and $\bar{\mathbf{\Theta}}$. The final output of the node feature at the R-th layer encodes the node's appearance information and structural properties within its R-hop neighbors, and can be used for the node classification, the link prediction and the node recommendation and so on.

The aggregation process in the node embedding stage allows the node to absorb the neighbors' information and then obtain its feature representation. Due to the information integration within the same cluster, the node's representation becomes more discriminative, which is beneficial to the subsequent classification task. However, the multi-layer aggregation (*i.e.*, the value R is set too high) may mix the features of all

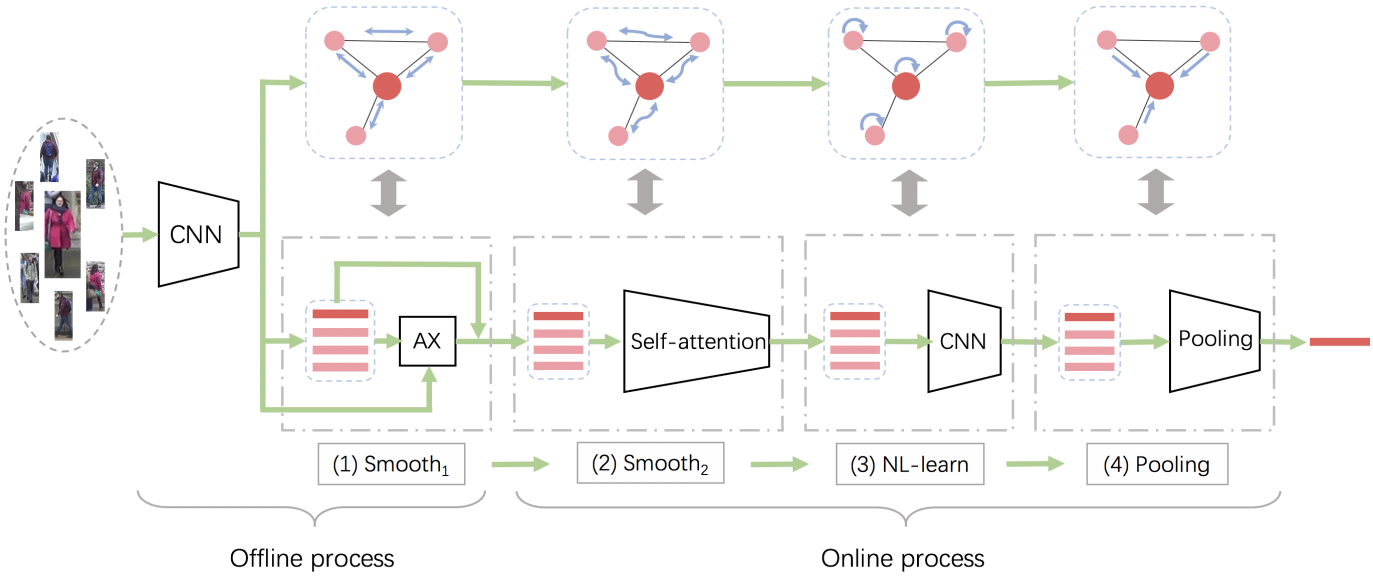

Fig. 2. The pipeline of the proposed progressive context-aware graph feature learning method. 'CNN', 'Self-attention' and 'Pooling' denote the convolutional neural network, the multi-head self-attention mechanism and the pooling operation, respectively. 'AX' represents the multiplication of the feature matrix $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$. 'Smooth$_1$', 'Smooth$_2$' and 'NL-learn' denote the first-order smoothing computation, the second-order smoothing computation and the non-linear mapping, respectively. The circle and rectangle represent the sample and its feature vector, respectively.

nodes in the graph and makes these features similar and indistinguishable, which is called the over-smoothing problem and possibly produces a negative effect on performance [27].

To make a further in-depth exploration for the GCN mechanism, we instantiates the node embedding stage in Eq. 1 by a classic update scheme [55]:

$$\mathbf{x}_i^{(r+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i} \tilde{\mathbf{A}}(i,j)\mathbf{x}_j^{(r)}\mathbf{W}^{(r)}\right), \quad r = 1, \cdots, R, \tag{2}$$

where $\tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\tilde{\mathbf{D}}^{-\frac{1}{2}}$ denotes the normalized adjacency matrix with $\mathbf{I}$ as an identity matrix and $\tilde{\mathbf{D}} = \sum_j \mathbf{A}(j,j)$, $\mathcal{N}_i$ is a set of the nodes connecting to the node i, $\mathbf{W}^{(r)}$ is the learnable transformation weight at the r-th layer, $\sigma(\cdot)$ is a non-linear activation function, such as ReLU. The node embedding in Eq. 2 can be re-written in matrix form:

$$\mathbf{X}_p^{(r+1)} = \sigma\left(\tilde{\mathbf{A}}\mathbf{X}^{(r)}\mathbf{W}^{(r)}\right), \quad r = 1, \cdots, R. \tag{3}$$

In each layer, the node feature's update is actually comprised of two sub-updates: firstly updating the node feature with the aid of contextual information by passing message from interconnected nodes ($\mathbf{X}'^{(r)} \leftarrow \tilde{\mathbf{A}}\mathbf{X}^{(r)}$), then further updating the node feature by learning a non-linear feature mapping ($\mathbf{X}^{(r+1)} \leftarrow \sigma\left(\mathbf{X}'^{(r)}_p\mathbf{W}^{(r)}\right)$). Essentially it is an alternating iterative process. The context-aware message passing process and the further feature non-linear mapping process alternate with each other to update the node feature representations.

In the graph-focused application, it concentrates on computing a graph feature representation, thus there is a graph embedding stage besides the node embedding stage. In the graph embedding stage, the graph-level feature representation is obtained by a pooling operation, and can be formatted as:

$$\mathbf{x}_{\mathcal{G}} = \gamma\left(\left\{\mathbf{x}_i^{(R)} | v_i \in \mathcal{G}\right\}\right), \tag{4}$$

where γ is a pooling function, such as the global sum pooling that sums the feature vectors over all nodes in the graph. The final graph feature representation $\mathbf{x}_{\mathcal{G}}$ summarizes all nodes' feature representations in the graph and encodes the graph topological structure, and can be used for representing the chemical structure of molecule compounds, classifying the graph, and so on.

Most GCN-based Re-ID methods [20], [42], [43], [23] can be categorized into the node-focused application and the resulting node feature representation is adopted as the final representation for Re-ID. By contrast, we construct a graph data to represent the sample and compute the graph feature representation as the sample representation, which belongs to the graph-focused GCN application.

B. Proposed Graph Feature Learning

Graph construction. Starting from the initial individual feature representations of all samples in the dataset, which can be readily obtained by most existing offline Re-ID baselines, we construct the graph $\mathcal{G}_p$ for a target sample p.

The graph is denoted as $\mathcal{G}_p = (V_p, E_p)$ with the node feature matrix $\mathbf{X_p}$ and the adjacency matrix $\mathbf{A_p}$, where V_p is a set of sample nodes and E_p is a set of edges.

To be specific, V_p is composed of the anchor sample p and its context samples in $\mathcal{C}_p$,

$$V_p = \{p\} \cup \mathcal{C}_p, \tag{5}$$

where $\mathcal{C}_p$ is a set of the global context samples of the sample p by searching its k-nearest neighbors (k-nn) in the whole

dataset in the initial feature space. Formally,

$$\mathcal{C}_p = N(p, K) = \{p_1, p_2, \cdots, p_K\}, \quad |N(p, K)| = K, \quad (6)$$

where $S(p, p_1) > S(p, p_2) > \cdots > S(p, p_K)$ and S denotes the similarity between samples' initial representations and is same as the similarity measure in the Re-ID baseline.

The set E_p stores the connection relations between nodes in V_p and is made up of the anchor-related edge set and the context-related edge set,

$$E_p = E(p \leftrightarrow \mathcal{C}_p) \cup E(\mathcal{C}_p \leftrightarrow \mathcal{C}_p), \quad (7)$$

where

$$\begin{aligned} E(p \leftrightarrow \mathcal{C}_p) &= \{e_{ij} | i \in \{p\}, j \in \mathcal{C}_p\}, \\ E(\mathcal{C}_p \leftrightarrow \mathcal{C}_p) &= \{e_{ij} | (i, j \in \mathcal{C}_p) \wedge (S(i, j) \geq \sigma_p)\}, \end{aligned} \quad (8)$$

and e_{ij} represents that the nodes i and j are connected to each other, $\sigma_p = \max_{\{i \in \{p\}, j \in \mathcal{C}_p\}} S(i, j)$ denotes the threshold value for valid connection. The node feature matrix $\mathbf{X}_p = \left[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n_p}\right] \in \mathbb{R}^{n_p \times d}$ stores the samples' d-dimensional initial individual feature vectors with $n_p = K+1$. The adjacency matrix $\mathbf{A}_p \in \mathbb{R}^{n_p \times n_p}$ digitizes the relations between nodes in E_p by:

$$\mathbf{A_p} = \begin{cases} S(i, j) & e_{ij} \in E_p \\ 0 & otherwise. \end{cases} \quad (9)$$

In general, based on a Re-ID baseline from which we obtain the initial individual feature representations and the similarity between them, the graph data for all samples in the dataset can be constructed offline.

For the feature learning of the graph $\mathcal{G}_p$, we propose a progressive context-aware graph feature learning method, in which there is a node embedding stage followed by a graph embedding stage. The network is illustrated in Fig.2.

Node embedding stage. From section III-A, GCNs is essentially a two-step alternating iterative process. Applying the classical GCN [55] in Eq. 2 to Re-ID can be interpreted as following: the message passing process makes the sample p's node representation absorb its contextual information in $\mathcal{C}_p$ and obtains its context-aware feature representation, and the non-linear mapping process updates the context-aware feature representation by a learnable non-linear transformation.

The message passing process plays a more important role than the non-linear mapping process in assembling contextual information into the sample p. From this, we deem that the message passing process is crucial for improving performance compared to the non-linear mapping process. We therefore propose a novel node embedding model by emphasizing the message passing process and simplifying the non-linear mapping process to achieve robust performance at a low cost.

Inspird by the method [26], we first remove the non-linear learning functions between each layers and only keep the final non-linear function in Eq. 3, and then derive the simplified node embedding model. The R-layer node feature matrix for the graph $\mathcal{G}_p$ is computed by:

$$\mathbf{X}_p^{(R)} = \sigma\left(\tilde{\mathbf{A}}_p^R \mathbf{X}_p \mathbf{W}\right), \quad (10)$$

where $\tilde{\mathbf{A}}_p^R$ is $\tilde{\mathbf{A}}_p$ to the power of R and $\mathbf{W} = \mathbf{W}^{(1)}\mathbf{W}^{(2)}\cdots\mathbf{W}^{(R)}$ is reparameterized into a single learnable transformation matrix. The simplified model in Eq. 10 can be interpreted that the message passing process ($\mathbf{X}_p' \leftarrow \tilde{\mathbf{A}}_p^R \mathbf{X}_p$) and the non-linear mapping process ($\mathbf{X}_p^{(R)} \leftarrow \sigma\left(\mathbf{X}_p' \mathbf{W}\right)$) are performed successively as two completely separated steps.

However, we know from section III-A that the repeated passing too many times could lead to the situation that the feature representations of nodes in the graph tend to be the same and the original information of nodes fades away gradually over the repeated passing. It could cause the over-smoothing problem and a further performance penalty. We therefore further optimize the message passing process.

We propose a progressive message passing process with the help of a fixed adjacency relationship among nodes followed by a learnable adjacency relationship. The fixed relationship is obtained based on the initial individual feature similarity measure from the Re-ID baseline, by which we compute the new feature representation of the node with the aid of information of the 1-hop neighbors. The learnable relationship is obtained based on a multi-head self-attention mechanism [56] in the updated feature space, by which we further compute the new representation with the help of information of the 2-hop neighbors.

Specifically, we first compute the node features by a first-order smoothing computation:

$$\dot{\mathbf{X}}_p = concat(\mathbf{X}_p, \mathbf{A}_p \mathbf{X}_p), \quad (11)$$

which $concat$ denotes the concatenation operation. We adopt the concatenated features as the node features for obtaining the feature representation with better expressive power. The first-order smoothing computation can be performed offline.

Then we apply the multi-head self-attention mechanism to further optimize the node features by a second-order smoothing computation:

$$\ddot{\mathbf{X}}_{(h)p} = softmax\left(\frac{(\dot{\mathbf{X}}_p \mathbf{W}_h^Q)(\dot{\mathbf{X}}_p \mathbf{W}_h^K)^\top}{\sqrt{2d}}\right)\left(\dot{\mathbf{X}}_p \mathbf{W}_h^V\right), \quad (12)$$

$$\ddot{\mathbf{X}}_p = concat(\ddot{\mathbf{X}}_{(1)p}, \cdots, \ddot{\mathbf{X}}_{(H)p})\mathbf{W}, \quad (13)$$

where $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$ and $\mathbf{W}_h^V$ are the $2d \times 2d$-dimensional learnable transformations at the h-th head ($h = 1, \cdots, H$), and $\mathbf{W} \in \mathbb{R}^{2hd \times 2d}$ is a learnable dimensionality reduced transformation. The H sub-adjacency relationships are implicitly learned by H heads from different perspectives and obtains H node feature matrices by Eq. 12. Then the H feature matrices are summarized to obtain the updated node features $\ddot{\mathbf{X}}_p$ in Eq. 13.

It can be seen that the proposed progressive message passing process is a dynamically self-adaptive process, in which we adopt different message passing modes in two different neighbor scales.

In the end, we add a two-layer non-linear mapping process and obtain the resulting node features:

$$\hat{\mathbf{X}}_p = \sigma\left(\sigma\left(\ddot{\mathbf{X}}_p \mathbf{W}_1\right)\mathbf{W_2}\right), \quad (14)$$

where $\mathbf{W}_1 \in \mathbb{R}^{2d \times d}$ is a learnable dimensionality reduced transformation and $\mathbf{W}_2 \in \mathbb{R}^{d \times d}$ is a learnable linear trans-

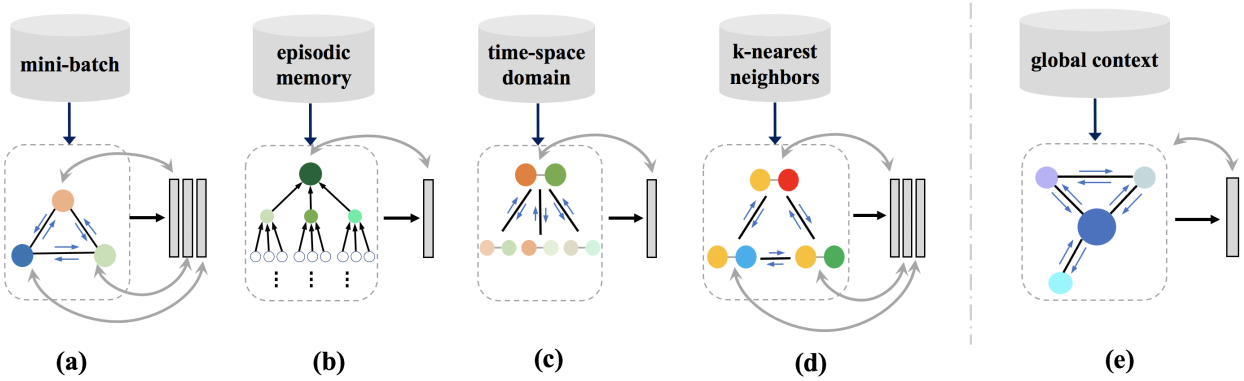

Fig. 3. Illustration of four categories of existing GCN-based target Re-ID methods in (a)-(d) and the proposed method in (e). The circle represents the sample and the solid circles with the same color denote the same sample. The rectangle represents the node feature representation.

formation, and we adopt ReLU as the non-linear activation function σ.

Graph embedding stage. The node feature representation $\hat{\mathbf{x}}_p$ in $\hat{\mathbf{X}}_p$ already contains the contextual information and can be used as the final feature representation of the sample p for target Re-ID. However, we hypothesize that the graph feature representation strengthens the graph's topological and structural information compared to the node representation. Thus the feature representation of the sample p is further explored by aggregating the feature representations of all nodes in the graph and computing the graph feature representation.

The final graph feature representation is given by:

$$\mathbf{x}_{\mathcal{G}_p} = \sum_{i \in V_p} \omega_i \hat{\mathbf{x}}_i, \quad \omega_i = softmax(\mathbf{w}\hat{\mathbf{x}}_i), \tag{15}$$

where $\mathbf{w} \in \mathbb{R}^{1\times d}$ is a learnable vector.

We obtain the final feature representation of the sample p as:

$$\mathbf{f}_p = concat(\mathbf{x}_p, \mathbf{x}_{\mathcal{G}_p}). \tag{16}$$

C. Discussion with Other Related Methods

We discuss the differences between the proposed method and other related GCN-based target Re-ID methods [24], [43], [54], [23], [20], [42], [2] in detail in this section. The related methods can be divided into four categories according to their design logic, which are shown in Fig. 3. We analyze the difference in the aspects of the graph construction and the method architecture.

For the first category [24], [43], [54] in Fig. 3 (a), the graph is constructed straightforwardly by using all samples in a mini-batch as the nodes, and the node features are iteratively updated and optimized by combining the individual feature extraction network with classic node-feature update network. The context samples are explored only in a random subspace (*i.e.*, a mini-batch), resulting in learning a suboptimal feature representation.

For the second category [23] in Fig. 3 (b), the context samples are updated in real-time by introducing an episodic memory module in the training process, in which the samples' feature representations are stored and updated in real-time. A tree graph is iteratively constructed by setting the anchor sample as the root node and its context as the leaf nodes, and the root node's feature representation is computed by continuously absorbing the information of the leaf nodes. It can be seen that the global reliable context samples are explored at the cost of high model complexity.

For the third category [20] in Fig. 3 (c), the sample's neighboring co-travelers in the time-space domain are explored as its context and the graph is constructed with a probe-gallery node pair and the context pairs. The relation representation of the probe-gallery pair is learned by the classic GCN for Re-ID.

For the fourth category [42], [2] in Fig. 3 (d), a probe sample and its k-nn gallery samples are used to construct the graph with node modeling the relation feature of the probe-gallery pair and edge modeling the similarity between gallery samples. The relation feature is optimized by means of the classic GCN for Re-ID. The motivation behind this graph relies on the similarity relation transitivity, *i.e.*, when a is similar with b and b is similar with c, a high similarity between a and c can be derived.

In contrast, the proposed method in Fig. 3 (e) represents the target sample by a graph structured data, which encodes the first-order appearance information of the sample with its global context and the second-order relationship information among them. The context samples are obtained by searching the neighbors in the whole data space at a low cost. We propose a novel progressive context-aware graph feature learning method based on an in-depth look at the GCN mechanism and its relation to the Re-ID problem rather than the simple application. The proposed method is a high-level exploration and utilization of contextual information, and can achieve performance enhancement with its compact and flexible architecture based on most existing offline Re-ID baselines.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate the proposed method on four person Re-ID datasets **CUHK03** [57], **Market1501** [17], **DukeMTMC** [58], **MSMT17** [59], and a vehicle Re-ID dataset **VeRi-776** [60].

TABLE I
STATISTICS OF TARGET RE-ID DATASETS. #T-IDS DENOTES THE NUMBER OF TARGET IDENTITY FOR TRAINING, #P-IDS/G-IDS DENOTE THE NUMBERS OF PROBE/GALLERY IDENTITIES FOR TESTING.

Dataset	#TIDs	#PIDs	#GIDs	#images	#cams
MSMT17	1041	3060	3060	126441	15
DukeMTMC	702	702	1110	36411	8
Market1501	751	750	751	32668	6
CUHK03	767	700	700	28192	2
VeRi-776	576	200	200	51035	20

MSMT17 contains 126,441 images which belong to 4,101 identities from 15 cameras. These images are divided into the training set including 32,621 images of 1,041 identities, and the test set including 93,820 images of 3,060 identities. Compared with other Re-ID datasets, it is a larger and more challenging dataset.

DukeMTMC is composed of 36,411 images of 1,404 identities and is collected by eight high-resolution cameras. According to the database setting, the training set consists of 16,522 images of 702 identities and the test set includes 2,228 query images and 17,661 gallery images of 702 identities.

Market-1501 is collected by six disjoint cameras at the Tsinghua University campus, and it consists of 32,668 images of 1,501 identities. According to the database setting, 12,936 images from 751 identities are utilized as the training set, while 3,368 probe images and 19,732 gallery images from the remaining 750 identities are used as the test set.

CUHK03 includes 14,097 images of 1,467 identities captured by six cameras, of which 7,365 images of 767 identities are used as the training set and 6,732 images of 700 identities are used as the test set. There are 1,400 query images and 5,332 gallery images in the test set. CUHK03 provides both hand-labeled and DPM-detected bounding boxes [61]. We present results on both 'labeled' and 'detected' settings in the experiments.

VeRi-776 is a large-scale urban surveillance vehicle dataset for vehicle Re-ID. It contains over 50,000 images of 776 vehicles labeled with rich attributes, *e.g.* types, colors, brands, license plate annotation and spatiotemporal relation annotation. In this dataset, 37,781 images of 576 vehicles are applied as a training set and 11,579 images of 200 vehicles are employed as a test set. The query set is composed of 1678 images from the test set.

The statistic details of the five datasets are summarized in Table I.

Implementation and Evaluation. The experiments are conducted on a TITANV GPU with 12 GB of memory. During training, there are only 50 epochs for CUHK03 dataset and 20 epochs for the rest of datasets. The batch size is 128 with each identity containing 4 image samples. We adopt the Adam optimizer [62] with a warmup strategy for the learning rate and the objective function is the weighted summation of triplet loss, cross entropy loss and center loss, similar to the settings in the method [63]. During testing, the batch size is 32 and the target Re-ID ranking is computed by sorting the similarity values between the samples' feature representations. There are two parameters in the proposed method: the number of neighbors K for constructing the graph in Eq.6 and the number of heads H in the multi-head self-attention mechanism for learning the adjacency relationship in Eq.13. We set $K = 4$ and $H = 1$ in the experiments. The proposed method is evaluted based on several state-of-the-art CNN-based Re-ID models as the baseline: StrongB [63], FastReID [64] and AGW [19]. Two conventional evaluation metrics are reported: cumulated matching characteristics (CMC) and mean average precision (mAP).

B. Performance Analysis

1) Parameter analysis: Fig. 4 shows the results of the proposed method with the change of two parameters K and H on DukeMTMC and Market1501 datasets. We adopt the StrongB method [63] as the baseline in follow-up experiments unless otherwise stated.

With the increase of K, the Rank-1 and mAP results present an earlier increasing trend and then tend to be stable on two datasets. On the one hand, a larger K means that more context samples are introduced for constructing the graph and more contextual information can be explored for computing the graph feature representation. The increasing trend of the results indicates that the contextual information is utilized efficiently by the proposed method to improve performance. On the other hand, the disturbance is introduced with too large value of K. The proposed method however shows its strong robustness against the introduction of disturbance when sequentially increasing K. It might be due to our second-order smoothing computation in which a reliable adjacency relationship among samples is learned by the self-attention mechanism and indicates the weak connections between the anchor sample and the disturbance samples.

With the increase of H, there is a relatively stable performance on two datasets. The multi-head mechanism allows that the adjacency relationship among nodes is learned from H different perspectives. However, the context-filtering rule of k-nn leads to a relatively simple relationship among nodes at the semantic level. As a result, the sub-adjacency relationships learned from different heads are essentially similar, contributing to the proposed method's robustness against the change of H.

2) Ablation study: There are three key stages in the proposed method: the node embedding stage and the graph embedding stage for learning the sample's graph feature representation, the concatenation stage for obtaining its final feautre representation.

We assess the contributions of these stages to the performance on DukeMTMC and Market1501 datasets. As shown in Table II, the introduction of the node embedding stage brings a remarkable positive impact on results (StrongB (baseline) vs. Ours (w/ N)), indicating the effectiveness of the node embedding stage; the further introduction of the graph embedding stage also leads to better results (Ours (w/ N) vs. Ours (w/ N+G)), showing the superiority of the graph representation compared to the node representation for Re-ID; in addition,

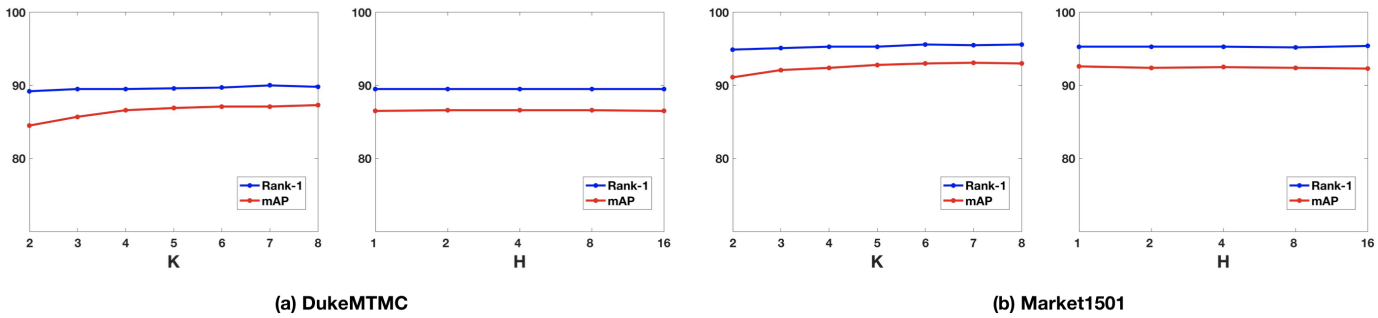

Fig. 4. The influence of the number of neighbors K for the graph construction and the number of heads H for the second-order smoothing computation on performance.

TABLE II
ABLATION STUDY OF THE PROPOSED METHOD. 'SMOOTH$_1$', 'SMOOTH$_2$', 'NL-LEARN', 'POOLING' AND 'CONCAT' ARE THE ABBREVIATIONS OF THE FIRST-ORDER SMOOTHING COMPUTATION IN EQ.11, THE SECOND-ORDER SMOOTHING COMPUTATION IN EQ.13, THE NON-LINEAR MAPPING PROCESS IN EQ.14, THE POOLING OPERATION IN EQ.15 AND THE CONCATENATION OF THE FEATURES IN EQ.16, RESPECTIVELY.

Models	Settings					DukeMTMC		Market1501	
	Node embedding stage			Graph embedding stage	Concat	Rank-1	mAP	Rank-1	mAP
	Smooth$_1$	Smooth$_2$	NL-learn	Pooling					
StrongB (baseline) [63]	×	×	×	×	×	86.4	76.4	94.5	85.9
Ours_N (w/o S$_1$)	×	✓	✓	✓	✓	88.1	86.3	94.6	92.3
Ours_N (w/o S$_2$)	✓	×	✓	✓	✓	88.3	86.6	94.3	92.6
Ours_N (w/o NL)	✓	✓	×	✓	✓	88.7	86.0	95.0	92.4
Ours (w/ N)	✓	✓	✓	×	×	88.9	85.8	95.5	91.7
Ours (w/ N+G)	✓	✓	✓	✓	×	89.3	86.9	95.4	92.5
Ours	✓	✓	✓	✓	✓	89.5	86.6	95.4	92.5

TABLE III
COMPARISON WITH CLASSICAL GCN METHODS ON DUKEMTMC AND MARKET1501 DATASETS.

Models		DukeMTMC		Market1501	
		Rank-1	mAP	Rank-1	mAP
StrongB (baseline) [63]		86.4	76.4	94.5	85.9
Node embedding stage	GCN-O [55]	88.3	82.6	95.1	88.7
	GIN [65]	88.3	81.8	95.0	88.1
	GraphSAGE [66]	87.8	81.7	94.8	87.8
	Ours (w/ N)	88.9	85.8	95.5	91.7
Graph embedding stage	GCN-O+G	88.2	82.4	95.2	88.7
	GIN+G	88.0	81.9	95.2	88.3
	GraphSAGE+G	87.9	81.9	95.1	88.0
	Ours	89.5	86.6	95.4	92.5

TABLE IV
INFLUENCE OF NUMBERS OF THE SMOOTHING COMPUTATION ON PERFORMANCE.

Iterations	DukeMTMC				Market1501			
	Smooth_1		Smooth_2		Smooth_1		Smooth_2	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	89.5	86.6	89.5	86.6	95.4	92.5	95.4	92.5
2	89.1	84.2	89.4	86.1	95.3	92.2	95.4	92.5
3	89.2	85.2	89.7	86.1	95.2	92.4	95.3	92.4
4	89.1	85.9	89.5	86.1	94.9	92.1	95.4	92.4

TABLE V
THE RUNNING TIME OF THE PROPOSED METHOD. THE OFFLINE TIME INCLUDES THE GRAPH CONSTRUCTION'S TIME AND THE FIRST-ORDER SMOOTHING COMPUTATION'S TIME.

Dataset	Number of images	Time (s)	
		Offline	Online
DukeMTMC	36,411	124.68	160.33
Market1501	32,668	110.49	130.26
CUHK03(detected)	14,096	27.57	166.78

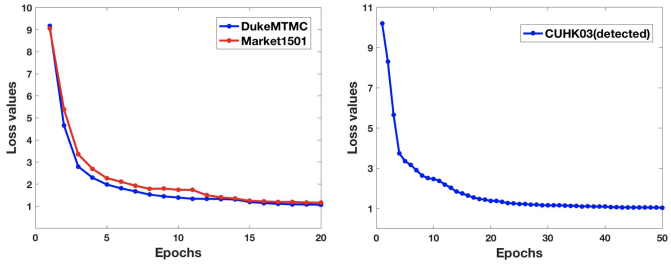

Fig. 5. Illustration for the convergence of the proposed method.

there is a similar performance for the proposed method with the graph representation and that with the concatenated representation (Ours (w/ N+G) vs. Ours). It may because the graph feature representation already carries the individual appearance information in the concatenated representation. To improve the proposed method's robustness, we adopt the concatenated feature for Re-ID.

We analyze the effectiveness of the successive operations in the node embedding stage. As shown in Table II, the higher Rank-1 accuracy is obtained by Ours than that by Ours_N (w/o S_1) or Ours_N (w/o S_2), indicating that both of the proposed first-order smoothing computation and the second-order smoothing computation bring the positive effects on performance, specifically, the first-order computation has a slight advantage over the second-order computation (Ours_N (w/o S_1) vs. Ours_N (w/o S_2)); in addition, the comparison between Ours_N (w/o NL) and Ours illustrates the effectiveness of the non-linear mapping process.

3) Comparison with classical GCN methods: The effectiveness of the proposed method has been proved in Section IV-B2. In this section, we verify that the effectiveness is mainly owing to the full exploitation and utilization of the contextual information by the proposed method. We adopt several classical GCN methods [55], [65], [66] instead of our graph feature learning method for Re-ID and report the comparison results in Table III[1].

Firstly, the performance is improved by applying either the proposed method or these GCN methods on the StrongB baseline both in the node embedding stage and in the further graph embedding stage, which provides a powerful support for the usefulness of the contextual information and the superiority of the graph representation for Re-ID. Secondly, it can be seen that the proposed method has an overwhelming advantage over these GCN methods in the node embedding stage, for example, Ours (w/ N) achieves 3.2% and 3.0% mAP gains compared to the state-of-the-art GCN-O [55] on two datasets, respectively. It indicates that our node embedding scheme fits better for the Re-ID task. Finally, based on the performance advantages in the node embedding stage, the proposed method shows the most significant performance improvement and also achieves the best results by further adding graph embedding stage compared to these GCN methods with our graph embedding stage[2], for instance on DukeMTMC dataset, 0.6% and 0.8% gains in the proposed method but 0.1% and 0.2% gains in the best GraphSAGE method at Rank-1 and mAP. It indicates the advancement of the overall framework in the proposed method.

4) The message passing vs. the non-linear mapping: Since the contextual information interaction occurs in the message passing process rather than the non-linear mapping process in GCNs, we assume that the message passing is more important than the non-linear mapping in improving Re-ID performance and derive the proposed method by strengthening the message passing and weakening the non-linear mapping. We verify the assumption by the experimental comparisons in this section. GCN-O+G in Table III can be viewed as the proposed method with the classic node embedding scheme by Eq. 2. Starting from this, we simplify the non-linear mapping and derive the proposed method with the node embedding scheme by Eq. 10, of which the results are presented by Ours_N (w/o S_2) in Table II and are better by a large margin than that of GCN-O+G at mAP. Based on Ours_N (w/o S_2), we enhance the

[1]We set the number of GCN layers to 1 for GCN-O [55] and GIN [65], and 2 for GraphSAGE [66] via 2-fold cross validation in DukeMTMC.

[2]These GCN methods are proposed with a focus on the node embedding stage for the node-focused application, thus there is no graph embedding stage in these methods.

TABLE VI
COMPARISON ON PERSON RE-ID DATASETS. WE RERUN THE PUBLISHED CODE OF THE METHOD MARKED WITH '*' ACCORDING TO THE AUTHORS' DEFAULT SETTING AND REPORT THE RESULTS. THE BEST RESULTS (EXCEPT FOR THE RE-RANKING RESULTS) ARE SHOWN IN RED.

Methods		MSMT17		DukeMTMC		Market1501		CUHK03 labeled		CUHK03 detected	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
BAT-net [67]	ICCV19	79.5	56.8	87.7	77.3	95.1	87.4	78.6	76.1	76.2	73.2
ABD-Net [68]	ICCV19	82.3	60.8	89.0	78.6	95.6	88.3	-	-	-	-
Pyramid-reid [5]	CVPR19	-	-	89.0	79.0	95.7	88.2	78.9	76.9	78.9	74.8
StrongB [63]	TMM19	72.5	48.3	86.4	76.4	94.5	85.9	63.8	61.2	58.3	58.0
StrongB(IBNa)* [63]	TMM19	77.4	54.2	89.0	78.8	95.0	87.5	65.6	63.8	62.3	60.6
CSPR-Net [33]	TMM19	75.3	50.8	83.5	71.9	94.2	84.8	64.7	62.8	-	-
FastReID* [64]	arXiv20	84.9	62.8	91.9	83.5	95.9	90.4	81.1	78.7	77.6	74.4
VA-reID [69]	AAAI20	-	-	91.6	84.5	96.2	91.7	-	-	-	-
CBN [70]	ECCV20	-	-	84.8	70.1	94.3	83.6	-	-	-	-
RGA-SC [8]	CVPR20	80.3	57.5	-	-	96.1	88.4	81.1	77.4	79.6	74.5
DF-HMC [38]	TMM20	74.3	43.6	83.3	68.2	93.8	81.8	-	-	-	-
3D-SF [71]	CVPR21	-	-	88.2	76.1	95.0	87.3	-	-	-	-
CDNet [72]	CVPR21	78.9	54.7	88.6	76.8	95.1	86.0	-	-	-	-
StrongB+FIDI [11]	TMM21	-	-	88.1	77.5	94.5	86.8	75.0	73.2	72.1	69.1
FA-Net [73]	TIP21	76.8	51.0	88.7	77.0	95.0	84.6	-	-	-	-
AGW* [19]	TPAMI21	74.4	49.0	88.5	79.3	95.2	88.5	73.3	72.5	70.6	70.0
SGGNN [42]	ECCV18	-	-	81.1	68.2	92.3	82.8	-	-	-	-
DFL-SGLE [18]	PR18	-	-	-	-	83.6	63.4	-	-	-	-
AFF-GNN [43]	ACMMM19	-	-	87.6	85.3	95.6	92.7	-	-	68.2	71.6
SFT [24]	ICCV19	73.6	47.6	86.9	73.2	93.4	82.7	68.2	62.4	-	-
MNE [23]	ICCV19	-	-	90.4	87.5	-	-	77.4	77.7	-	-
CAGCN [2]	AAAI21	-	-	91.3	85.9	95.9	91.7	-	-	-	-
Ours+StrongB		78.1	63.8	89.5	86.6	95.4	92.5	73.0	76.9	70.7	74.9
Ours+StrongB(IBNa)*		82.3	69.0	91.4	87.2	95.7	93.1	75.6	79.3	72.1	75.6
Ours+AGW*		80.2	60.5	90.2	84.3	96.3	91.9	79.4	80.1	77.6	77.8
Ours+FastReID*		**87.2**	**70.5**	**92.8**	**87.4**	**96.4**	**93.2**	**84.9**	**85.2**	**83.1**	**82.3**
Ours+FastReID*+rerank [12]		-	-	93.6	90.1	96.7	94.6	88.6	90.2	86.6	87.7

message passing by adding a trainable message passing and derive the proposed method (Ours), which achieves a further improvement at Rank-1 compared to the Ours_N (w/o S_2) in Table II. Overall, the performance is continuously improved when strengthening the message passing and weakening the non-linear mapping.

5) Analysis of over-smoothing: The multiple message passing processes could lead to the over-smoothing problem. We conduct the experiments by increasingly iterating the first-order smoothing computation or the second-order smoothing computation in our proposed progressive message passing process, and verify its effectiveness for mitigating the over-smoothing problem.

As shown in Table IV, the over-smoothing problem caused by the multiple first-order smoothing computations can be observed with a declining trend in performance with increased of iteration times. By contrast, there is no over-smoothing problem in the second-order smoothing computation, and almost stable performances are obtained with different iterations of the second-order computation. The adjacency relationship among nodes is learned self-adaptively in the second-order computation, and such relationship dominates a self-motivated message passing process which is then immune to the over-smoothing problem. Our progressive message passing process in which Iterations=1 is set for both of Smooth_1 and Smooth_2 obtains the best performance in Table IV. It shows the availability of our proposed process for mitigating the over-smoothing problem.

6) Efficiency analysis: Fig. 5 and Table V show the iterative convergence process in the training phase and the running time of the proposed method on DukeMTMC, Market1501 and CUHK03(detected) datasets, respectively. It can be seen that the proposed method converges rapidly and takes a short time for the online training accordingly. The time of the offline process is also acceptable.

TABLE VII
COMPARISON ON A VEHICLE RE-ID DATASET. WE RERUN THE PUBLISHED CODE OF THE METHOD MARKED WITH '*' ACCORDING TO THE AUTHORS' DEFAULT SETTING AND REPORT THE RESULTS. THE BEST RESULTS (EXCEPT FOR THE RE-RANKING RESULTS) ARE SHOWN IN RED.

Methods		VeRi-776	
		rank-1	mAP
VANet [74]	ICCV19	89.8	66.3
PrNd [75]	CVPR19	94.3	74.3
StrongB* [63]	TMM19	96.4	79.8
FastReID* [64]	arXiv20	96.9	81.3
CFVMNet [76]	ACMMM20	95.3	77.1
SPAN+CPDM [77]	ECCV20	94.0	68.9
SAVER [78]	ECCV20	96.4	79.6
PVEN [79]	CVPR20	95.6	79.5
StrongB+FIDI [11]	TMM21	95.7	77.6
AGW* [19]	TPAMI21	94.5	75.2
PCRNet [46]	ACMMM20	95.4	78.6
SGAT [80]	ACMMM20	89.7	65.7
CAGCN [2]	AAAI21	95.8	79.6
Ours+StrongB*		96.7	82.9
Ours+AGW*		94.6	75.6
Ours+FastReID*		**97.0**	**82.8**
Ours+FastReID*+rerank [12]		97.3	85.3

C. Comparison with State-of-the-arts

In this section, we compare the proposed method with the state-of-the-art individual-based and context-based methods on four person Re-ID datasets and a vehicle Re-ID dataset. The results are reported in Table VI and Table VII.

In comparison with the individual-based methods, the proposed method with different baselines achieves competitive performance on all datasets. Specifically, the proposed method with the FastReID as baseline [64] performs the best results with all comparisons. It is worth noting that the proposed method could adopt any individual-based methods as the baseline for further enhancement and then can readily remain the state-of-the-art performance with the robust baseline.

In comparison with the context-based methods, the proposed method surpasses all of them on all datasets. The architectures of these context-based methods are designed with sophisticated network structures and based on the joint optimization of individual-based feature representation and context-based feature representation. However, the architecture of the proposed method allows flexible selection of the individual-based baseline for the context-based feature representation optimization. We also report the re-ranking results of the proposed method by utilizing the PBCM post-processing method [12] and the performance boosting is obtained on all dataset[3].

In addition, Fig. 6 visualizes the distance relations between the samples' feature vectors across views from the StrongB baseline [63] and the proposed method. We randomly select seven identities with their images for computing the distance matrix. Seven small rectangles in a diagonal of the distance matrix represent the distance submatrices within each identity, respectively. The remaining region of the distance matrix describes the distance relations between different identities. It can be seen from Fig. 6 that for the distance matrix from the proposed method, the values of distance submatrices are closer to zero and others tend to be larger than the baseline, indicating that the proposed method effectively facilitates intra-class compactness and inter-class separability.

[3]Due to limited computer memory, the PBCM post-processing method [12] can not be performed on MSMT17.

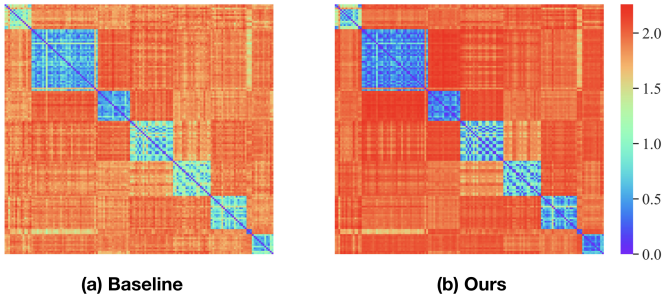

Fig. 6. Visualization of the distance matrix from the StrongB baseline [63] and the proposed method on DukeMTMC. We randomly select seven identities with their images and show the distance matrix between the features of images across views by arranging their identities in turn.

V. CONCLUSION

In this paper, we propose to learn a discriminative feature embedding with the aid of graph-based contextual information for target Re-ID. We adopt an individual feature extractor as the baseline, the global context of each sample is firstly explored in the whole dataset to construct the graph. For graph feature learning, we analyze the nature of the GCN mechanism applied in target Re-ID, and propose a progressive context-aware graph feature learning method. The learned graph feature representation conveys the first-order appearance information of the sample and its context and the second-order topological relationship among them, and is concatenated into the final representation of the sample. The proposed method enjoys simplicity and effectiveness, which are fully verified by extensive ablations and compared experiments on four widely-used person Re-ID datasets and one vehicle Re-ID dataset.

REFERENCES

[1] Y. Ge, D. Chen, F. Zhu, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," *arXiv preprint arXiv:2006.02713*, 2020.
[2] D. Ji, H. Wang, H. Hu, W. Gan, W. Wu, and J. Yan, "Context-aware graph convolution network for target re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
[3] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, and S. Satoh, "Beyond intra-modality: A survey of heterogeneous person re-identification," in *IJCAI*, 2020.
[4] J. Deng, M. S. Khokhar, M. U. Aftab, J. Cai, R. Kumar, J. Kumar *et al.*, "Trends in vehicle re-identification past, present, and future: A comprehensive review," *arXiv preprint arXiv:2102.09744*, 2021.
[5] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8514–8522.

[6] 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 International Conference on Computer Vision*, 2019, pp. 3642–3651.

[7] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 667–676.

[8] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.

[9] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3691–3701.

[10] M. Cao, C. Chen, X. Hu, and S. Peng, "Towards fast and kernelized orthogonal discriminant analysis on person re-identification," *Pattern Recognition*, vol. 94, pp. 218–229, 2019.

[11] C. Yan, G. Pang, X. Bai, C. Liu, N. Xin, L. Gu, and J. Zhou, "Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss," *IEEE Transactions on Multimedia*, 2021.

[12] M. Cao, C. Chen, H. Dou, X. Hu, S. Peng, and A. Kuijper, "Progressive bilateral-context driven model for post-processing person re-identification," *IEEE Transactions on Multimedia*, 2020.

[13] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] J. Lu, W. Zhang, and H. Yin, "Generate and purify: Efficient person data generation for re-identification," *IEEE Transactions on Multimedia*, 2021.

[15] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 608–617.

[16] 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.

[17] 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.

[18] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, and N. Zheng, "Deep feature learning via structured graph laplacian embedding for person re-identification," *Pattern Recognition*, vol. 82, pp. 94–104, 2018.

[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[20] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2158–2167.

[21] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3289–3299.

[22] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

[23] S. Li, D. Chen, B. Liu, N. Yu, and R. Zhao, "Memory-based neighbourhood embedding for visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6102–6111.

[24] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4976–4985.

[25] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2530–2539.

[26] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International conference on machine learning*. PMLR, 2019, pp. 6861–6871.

[27] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[28] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[29] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[30] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5794–5803.

[31] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3750–3759.

[32] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[33] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua, "Concentrated local part discovery with fine-grained part representation for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1605–1618, 2019.

[34] Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian, and X. Zhou, "Adaptive graph representation learning for video person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 8821–8830, 2020.

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

[36] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[37] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2860–2871, 2019.

[38] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, and D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3180–3195, 2020.

[39] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7103–7112.

[40] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao, "Learning multi-granular hypergraphs for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2899–2908.

[41] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.

[42] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 486–504.

[43] Y. Li, H. Yao, L. Duan, H. Yao, and C. Xu, "Adaptive feature fusion via graph neural network for person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2115–2123.

[44] M. Cao, C. Chen, X. Hu, and S. Peng, "From groups to co-traveler sets: Pair matching based person re-identification framework," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2573–2582.

[45] H. Liu, Z. Xiao, B. Fan, H. Zeng, Y. Zhang, and G. Jiang, "Prgcn: Probability prediction with graph convolutional network for person re-identification," *Neurocomputing*, vol. 423, pp. 57–70, 2021.

[46] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 907–915.

[47] A. M. N. Taufique and A. Savakis, "Labnet: Local graph aggregation network with class balanced loss for vehicle re-identification," *arXiv preprint arXiv:2011.14417*, 2020.

[48] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[49] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[50] F. Hu, Y. Zhu, S. Wu, W. Huang, L. Wang, and T. Tan, "Graphair: Graph representation learning with neighborhood aggregation and interaction," *Pattern Recognition*, vol. 112, p. 107745, 2021.

[51] L. Cai, J. Li, J. Wang, and S. Ji, "Line graph neural networks for link prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] F. Errica, M. Podda, D. Bacciu, and A. Micheli, "A fair comparison of graph neural networks for graph classification," *International Conference on Learning Representations (ICLR)*, 2020.

[53] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun, and W. Wang, "Simgnn: A neural network approach to fast graph similarity computation," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 384–392.

[54] L. Bao, B. Ma, H. Chang, and X. Chen, "Masked graph attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[55] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[57] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[58] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[59] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[60] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[61] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[63] 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*, vol. 22, no. 10, pp. 2597–2609, 2019.

[64] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, vol. 6, no. 7, p. 8, 2020.

[65] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *International Conference on Learning Representations (ICLR)*, 2019.

[66] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[67] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8030–8039.

[68] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8351–8361.

[69] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, X. Sun, and W. Zheng, "Aware loss with angular regularization for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 114–13 121.

[70] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, and Q. Tian, "Rethinking the distribution gap of person re-identification with camera-based batch normalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 140–157.

[71] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, and W.-S. Zheng, "Learning 3d shape feature for texture-insensitive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8146–8155.

[72] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6729–6738.

[73] Y. Liu, W. Zhou, J. Liu, G.-J. Qi, Q. Tian, and H. Li, "An end-to-end foreground-aware network for person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2060–2071, 2021.

[74] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[75] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3997–4005.

[76] Z. Sun, X. Nie, X. Xi, and Y. Yin, "Cfvmnet: A multi-branch network for vehicle re-identification based on common field of view," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3523–3531.

[77] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," in *European Conference on Computer Vision*. Springer, 2020, pp. 330–346.

[78] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 369–386.

[79] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[80] Y. Zhu, Z.-J. Zha, T. Zhang, J. Liu, and J. Luo, "A structured graph attention network for vehicle re-identification," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 646–654.