Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Other Tokens Matter: Exploring Global and Local Features of Vision Transformers for Object Re-Identification

Yingquan Wang[a], Pingping Zhang[b,**], Dong Wang[a], Huchuan Lu[a]

[a]School of Information and Communication Engineering, Dalian University of Technology
[b]School of Future Technology, School of Artificial Intelligence, Dalian University of Technology

ABSTRACT

Object Re-Identification (Re-ID) aims to identify and retrieve specific objects from images captured at different places and times. Recently, object Re-ID has achieved great success with the advances of Vision Transformers (ViT). However, the effects of the global-local relation have not been fully explored in Transformers for object Re-ID. In this work, we first explore the influence of global and local features of ViT and then further propose a novel Global-Local Transformer (GLTrans) for high--performance object Re-ID. We find that the features from last few layers of ViT already have a strong representational ability, and the global and local information can mutually enhance each other. Based on this fact, we propose a Global Aggregation Encoder (GAE) to utilize the class tokens of the last few Transformer layers and learn comprehensive global features effectively. Meanwhile, we propose the Local Multi-layer Fusion (LMF) which leverages both the global cues from GAE and multi-layer patch tokens to explore the discriminative local representations. Extensive experiments demonstrate that our proposed method achieves superior performance on four object Re-ID benchmarks.

1. Introduction

Object Re-Identification (Re-ID) aims to retrieve specific objects from images taken at different times and places. It has drawn lots of attention due to its many real-world applications, such as safe communities, intelligent surveillance and criminal investigations (Ye et al., 2021; Zheng et al., 2016; Weng et al., 2023). Benefiting from the local modeling ability of Convolutional Neural Networks (CNNs) (He et al., 2016), CNN-based methods have dominated the object Re-ID (Sun et al., 2018; Wang et al., 2018a; Khatun et al., 2020; Lin et al., 2023; Zhang et al., 2023) over the past two decades. Among them, one of most straightforward and efficient strategies is to split the feature maps to obtain fine-grained cues (Sun et al., 2018; Wang et al., 2018a). The main idea is shown in Fig. 1(a). Although excellent performances are achieved, this kind of methods are limited by the poor global representation ability of the convolution operations (Peng et al., 2021).

Recently, Transformers as powerful structures (Vaswani

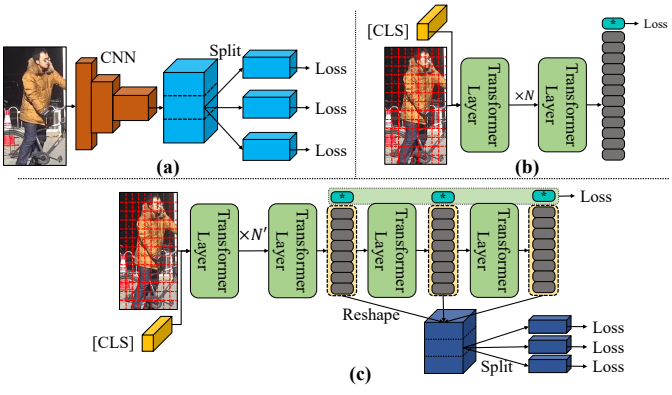

Fig. 1. Different structures employed in object Re-ID. (a) Part-based CNNs for local features. (b) Pure Transformers for global features. (c) Our proposed GLTrans method considers both local and global features.

**Corresponding author: Tel.: +0-000-000-0000; fax: +0-000-000-0000;
e-mail: zhpp@dlut.edu.cn (Pingping Zhang)

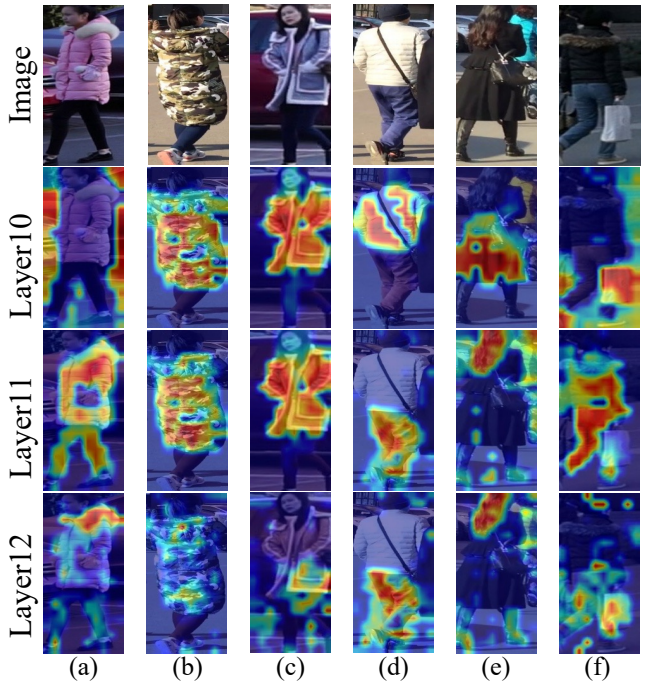

Fig. 2. Heatmap visualization of ViT's different layers by Gram-Cam (Selvaraju et al., 2017) on MSMT17. Specifically, Layer10, Layer11 and Layer12 mean the heatmap of the 10-*th*, 11-*th* and 12-*th* layers from ViT. Deeper red colors signify higher weights.

et al., 2017) have demonstrated superior performance for many visual tasks such as image classification and object detection (Vaswani et al., 2017; Dosovitskiy et al., 2021). The key reason is that Transformers aggregate information based on multi-head self-attention and focus on long-distance dependencies (Peng et al., 2021). Inspired by this fact, He *et al.* (He et al., 2021) introduce the first pure Transformer-based method for object Re-ID. Following that, many pure Transformer-based methods have been proposed (Lai et al., 2021a; Li et al., 2021c; Zhu et al., 2022; Liu et al., 2023b). As illustrated in Fig. 1 (b), the class token is employed to represent the entire image. However, these methods usually neglect two key issues: 1) The patch tokens contain rich fine-grained cues. 2) The features from the last few layers also have strong representations. To address these issues, some researchers divide the patch tokens into several independent regions for mining local discriminative cues. For instance, Zhu *et al.* (Zhu et al., 2021) employ the optimal transport algorithm (Cuturi, 2013) to discover local tokens with shared semantics and subsequently extract fine-grained information. Wang *et al.* (Wang et al., 2022b) and Zang *et al.* (Zang et al., 2022) reorganize patch tokens into feature maps and split these feature maps horizontally to extract local-wise representations. However, different from the feature map of CNNs, each patch token contains diverse global-view information, which needs to be further selected and refined. In addition, simple partitions may miss the structure information.

On the other hand, Fig. 2 visualizes the last three layers of ViT fine-tuned on the MSMT17 dataset (Wei et al., 2018). It is apparent that each layer emphasizes different semantics, and the patch tokens show strong representations due to the excellent global modeling of Transformers. For example, as depicted in Fig. 2 (d), the patch tokens of Layer-10 focus on the upper white coat, while that of Layer-11 and Layer-12 focus on the blue pants and shoes. Similar observations can be found in Fig. 2 (b) and Fig. 2 (c). In addition, we find that the most effective representations of patch tokens are not consistently found in the final layer. For instance, as shown in Fig. 2 (b), Fig. 2 (c) and Fig. 2 (f), the last layer is inferior to the shallower layers. Thus, using solely the features from the last layer may be a suboptimal choice. Furthermore, obtaining comprehensive fine-grained representations requires a full consideration of both the patch tokens and multi-stage features.

To this end, we propose a novel framework named Global-Local Transformer (GLTrans) to obtain a more robust and compact representation for object Re-ID. The main structure is shown in Fig. 1 (c), which is very different from previous works. More specifically, we first obtain the multi-layer patch tokens and class tokens from ViT (Dosovitskiy et al., 2021). Then, they are passed through a Local Multi-layer Fusion (LMF) and a Global Aggregation Encoder (GAE) for generating more discriminative local and global features, respectively. In GAE, the multi-layer class tokens are passed through a Fully-Connected (FC) layer for a comprehensive global feature. On the other hand, the multi-layer patch tokens are fed into the LMF to mine fine-grained information. Specifically, the LMF contains three parts: Patch Token Fusion (PTF), Global-guided Multi-head Attention (GMA) and Part-aware Transformer Layer (PTL). Firstly, the PTF re-weights and fuses the multi-layer patch tokens. Then the GMA further enhances patch tokens guided by global features. Finally, the PTL generates discriminative local features from multiple regions of enhanced patch tokens. Extensive experiments on four large-scale object Re-ID benchmarks demonstrate that our method shows better results than most state-of-the-art methods.

The main contributions are summarized as follows:

- We propose a novel learning framework (*i.e.*, GLTrans) to take local and global advantages of vision Transformers for robust object Re-ID.
- We propose the LMF to fuse multi-layer patch tokens for discriminative local representations. Additionally, we also present the GAE to aggregate multi-layer class tokens for comprehensive global representations.
- Extensive experiments demonstrate that our framework can effectively extract comprehensive feature representations. It achieves outstanding performances on four large-scale object Re-ID benchmarks.

2. Related Work

2.1. CNNs for Object Re-Identification

Recently, object Re-ID tasks have achieved great promotion in performance. Extracting discriminative cues is critical for object Re-ID. In the past few years, CNN-based methods have predominated the Re-ID tasks (Sun et al., 2018; Wang et al., 2018c; Li et al., 2018; Suh et al., 2018; Chen et al., 2019;

Huang et al., 2023). Early works extract robust features by applying deep CNNs. For example, Wang *et al.* (Wang et al., 2018c) cascade multiple convolutional layers to obtain different semantic information. However, these methods are time-consuming. On the other hand, some works (Sun et al., 2018; Wang et al., 2018a; Zang et al., 2022; Zhang et al., 2017; Sun et al., 2019b; Zheng et al., 2019a) split the feature maps into multiple parts along different directions for fine-grained information. Sun *et al.* (Sun et al., 2018) utilize horizontal partitions to learn discriminative local features. Zhang *et al.* (Zhang et al., 2017) not only partition the feature maps for learning local cues but also design a dynamic alignment mechanism to measure the similarity between the same semantic parts. Furthermore, Zheng *et al.* (Zheng et al., 2019a) divide deep feature maps into multi-scale sub-maps to incorporate local and global information. However, these methods treat every part equally, which may lose some crucial information. To address this issue, several attention-based methods (Si et al., 2018; Xu et al., 2018; Ye et al., 2024) are employed to suppress irrelevant features and enhance discriminative ones. In addition, Huang *et al.* (Huang et al., 2023) design the graph attention mechanism to extract useful person representations. Xu *et al.* (Xu et al., 2018) introduce pose estimation to guide attention generation. Inspired by (Wang et al., 2018b), Zhang *et al.* (Zhang et al., 2020b) generate attention maps by considering the pixel relation. Although these methods have made great progresses in object Re-ID, CNN-based features generally focus on the local discriminative regions, which may cause over-fitting and ignore global important information. To address these issues, in this work we introduce a novel pure Transformer framework for a more compact representation.

2.2. Transformers for Object Re-Identification

Due to its global modeling capabilities, Transformers become the mainstream models in the field of Natural Language Processing (NLP) (Floridi and Chiriatti, 2020). For vision tasks, Dosovitskiy *et al.* (Dosovitskiy et al., 2021) propose a ViT model and achieve excellent performances. Following ViT, some researchers (He et al., 2021; Zhang et al., 2021b; Lai et al., 2021b; Chen et al., 2021, 2022) introduce Transformers into the Re-ID field for robust and discriminative global representations. For example, He *et al.* (He et al., 2021) propose a pure Transformer model with the side information embeddings and designing a jigsaw patch module. Zhang *et al.* (Zhang et al., 2021b) introduce Transformer layers to hierarchically aggregate the multi-scale features from CNN-based backbones. Gao *et al.* (Gao et al., 2024) introduce part representation learning with a teacher-student decoder for occluded person Re-ID. Lu *et al.* (Lu et al., 2023) propose a progressive modality-shared Transformer for effective visible-infrared person Re-ID. Furthermore, Chen *et al.* (Chen et al., 2022) utilize the key points and Transformers to extract the structure-aware information for visible-infrared person Re-ID. However, these Transformer-based methods treat the class token as the feature representation, which may neglect the fine-grained cues among the patch tokens. To address this issue, Zhu *et al.* (Zhu et al., 2021) employ the optimal transport algorithm (Cuturi, 2013) to define the semantics of patch tokens and use local Transformers for local-aware representations. However, the semantics defined by the auxiliary algorithms may interfere with the complex background and mislead the feature representations. On the other hand, Wang *et al.* (Wang et al., 2022b) split the feature maps horizontally and align the patch tokens with the pose information for better local features. Although good performances are achieved, the introduced key-point algorithm (Sun et al., 2019a) is not reliable in many complex scenes, and they may neglect the spatial relation among the patch tokens. Recently, Yan *et al.* (Yan et al., 2023) integrate the advantages of CNNs and Transformers, and propose a convolutional multi-level Transformer for local-aware object Re-ID. Wang *et al.* (Wang et al., 2024) propose the token permutation for multi-spectral object Re-ID. Zhang *et al.* (Zhang et al., 2024) select diverse tokens from Transformers for multi-modal object Re-ID. Furthermore, Liu *et al.* (Liu et al., 2021) propose a trigeminal Transformers for video-based person Re-ID. Liu *et al.* (Liu et al., 2023a) propose deeply coupled convolution-Transformer with spatial–temporal complementary learning for video-based person Re-ID. Inspired by the great ability of vision-language models, Yu *et al.* (Yu et al., 2024) propose a text-free CLIP model for video-based person Re-ID. Different from previous works, we propose a local multi-layer fusion module to further mine the spatial information and directly extract the fine-grained cues from patch tokens.

3. Proposed Method

As illustrated in Fig. 3, the proposed framework (i.e., GLTrans) mainly includes three key modules: Vision Transformer (ViT), Global Aggregation Encoder (GAE) and Local Multi-layer Fusion (LMF). We will elaborate on these key components in the following subsections.

3.1. Revisiting Vision Transformer

To begin with, we briefly review the ViT (Dosovitskiy et al., 2021). Given an image $x \in \mathbb{R}^{H\times W\times C}$, where H, W and C represent the height, width and channel, respectively, we utilize a sliding window to obtain overlapped image patches. Then, the image is disassembled into N patches ($N = \lfloor\frac{H+S-P}{S}\rfloor\times\lfloor\frac{W+S-P}{S}\rfloor$). Specifically, we set the step size as $S = 12$ and the patch size as $P = 16$. Each patch is linearly projected into a D-dimensional vector. In addition, a class token v_1 is prefixed to these verctors. As a result, the vector sequence is obtained by concatenating these vectors, as follows:

$$\mathcal{X}_1 = [v_1; \phi(x_1^1); \cdots; \phi(x_1^N)] + \mathcal{P} + \mathcal{S}, \tag{1}$$

where $\mathcal{P} \in \mathbb{R}^{(N+1)\times D}$ and $\mathcal{S} \in \mathbb{R}^{(N+1)\times D}$ are learnable embeddings, representing the position information and side information (cameras or viewpoints), respectively.

3.2. Global Aggregation Encoder

Many Transformer-based methods (Dosovitskiy et al., 2021; Zhu et al., 2022) have shown that using the class token from the last layer achieves attractive results. However, as depicted in

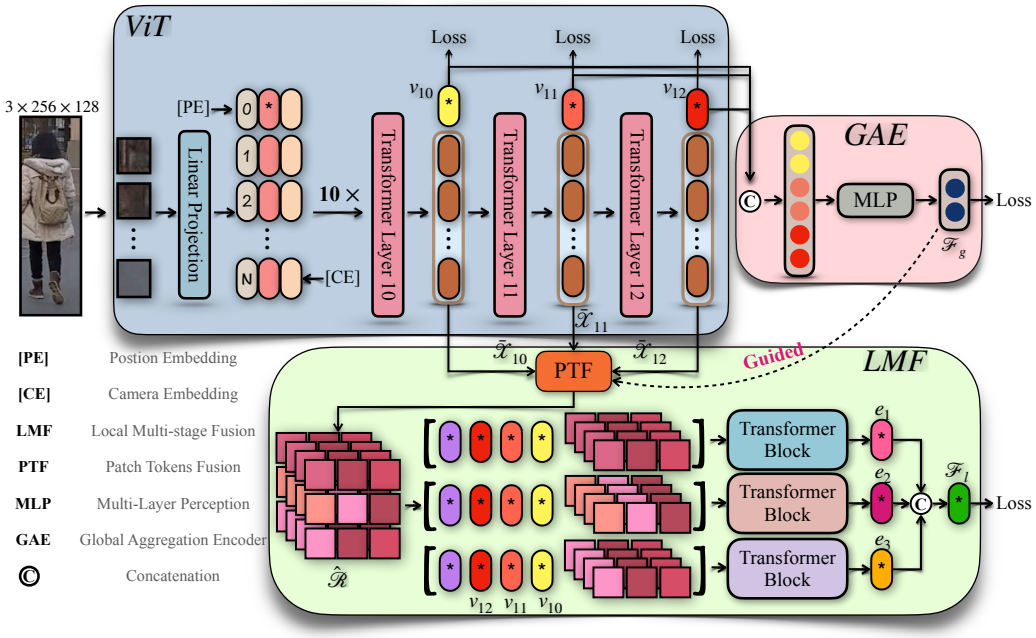

Fig. 3. Our proposed GLTrans. The Vision Transformer (ViT) with side information embedding (cameras or viewpoints) is adopted as the backbone to obtain multi-layer class tokens and patch tokens. Then, two branches are used to extract global and local representations. The Global Aggregation Encoder (GAE) generates global representations by incorporating multi-layer class tokens, while Local Multi-layer Fusion (LMF) takes patch tokens as inputs to further extract the local-wise discriminative features.

Fig. 2, the last few layers of the ViT also present strong representational abilities and the best representations are not always in the last layer. Therefore, utilizing the last class token may be a suboptimal strategy. To address this, we propose a GAE module. Specifically, we collect class tokens from the last three layers to construct a global aggregated vector $\hat{\mathcal{F}}_g$,

$$\hat{\mathcal{F}}_g = [v_{10}, v_{11}, v_{12}]. \tag{2}$$

For generating the useful global information and reducing the irrelevant cues, we directly feed the vector $\hat{\mathcal{F}}_g$ into a Fully-Connected layer (FC) and a GeLU activation function (Hendrycks and Gimpel, 2017) to obtain a comprehensive global representations,

$$\mathcal{F}_g = GeLU(FC(\hat{\mathcal{F}}_g)). \tag{3}$$

As shown in Fig. 2, the features from the last few layers exhibit strong discriminative and diverse semantics. However, simply concatenating or adding these features may enhance some erroneous information, as depicted in Fig. 2 (a) and Fig. 2 (f). Therefore, we introduce a FC layer to recognize the relationships among them and further highlight the useful ones. The experiments in Sec. 4.4 demonstrate that this simple structure can efficiently extract comprehensive features.

3.3. Local Multi-layer Fusion

There are some pure Transformer-based Re-ID methods to explore global and local representations (Song et al., 2023; Zhu et al., 2021). However, most of them extract the global and local representations independently, which may neglect that the local-global interaction can complement each other. On the other hand, most of previous methods (Dosovitskiy et al., 2021; He et al., 2021) adopt patch tokens from the last layer, which inevitably neglect diverse semantics from different layers. To address these issues, we propose the Local Multi-layer Fusion (LMF), which contains Patch Token Fusion (PTF), Global-guided Multi-head Attention (GMA) and Part-based Transformer Layers (PTL). They are described as follows.

3.3.1. Patch Token Fusion

As shown in Fig. 2, features from different Transformer layers have diverse semantics. They can complement each other. In addition, the cascade Transformer blocks neglect the spatial information among the patch tokens. Thus, we propose the PTF to obtain a compact local representation by aggregating multi-layer patch tokens and enhancing these spatial relationships.

Specifically, as shown in Fig. 4, we take the intermediate features $\mathcal{X}_{10}, \mathcal{X}_{11}, \mathcal{X}_{12}$ and split them into patch tokens $\tilde{\mathcal{X}}_l \in \mathbb{R}^{N\times D}$ and class token $v_l \in \mathbb{R}^{1\times D}$. Then, the patch tokens $\tilde{\mathcal{X}}_l$ are fed into two linear transformations followed by Sigmoid and GeLU activation functions to generate the weight mask $\mathcal{S}_l$,

$$\mathcal{S}_l = Sigmoid\left(GeLU(\tilde{\mathcal{X}}_l \times \mathcal{W}_l^1) \times \mathcal{W}_l^2\right), \tag{4}$$

where $\mathcal{W}_l^1 \in \mathbb{R}^{D\times \frac{D}{r_1}}$ and $\mathcal{W}_l^2 \in \mathbb{R}^{\frac{D}{r_1}\times D}$ are learnable parameters. r_1 is a reduction ratio of feature dimensions. Furthermore, we obtain enhanced patch tokens $\bar{\mathcal{X}}_l$ by

$$\bar{\mathcal{X}}_l = \mathcal{S}_l \otimes \tilde{\mathcal{X}}_l + \tilde{\mathcal{X}}_l, \tag{5}$$

where $\mathcal{S}_l \in \mathbb{R}^{N\times D}$. Afterwards, we reshape enhanced patch tokens $\bar{\mathcal{X}}_l$ according their spatial position and concatenate them,

$$\bar{\mathcal{X}} = Concat[\bar{\mathcal{X}}_{10}; \bar{\mathcal{X}}_{11}; \bar{\mathcal{X}}_{12}], \tag{6}$$

where $\bar{\mathcal{X}} \in \mathbb{R}^{\hat{H}\times\hat{W}\times 3D} (N = \hat{H}\times\hat{W})$. Inspired by previous works (Song et al., 2023; Peng et al., 2021), we also apply 1×1 and 3×3 convolutions following a BN and GeLU activation function to explore the spatial relation among the patch tokens,

$$\mathcal{R} = GeLU\left(BN\left(Conv(\bar{\mathcal{X}})\right)\right), \tag{7}$$

where $\mathcal{R} \in \mathbb{R}^{\hat{H}\times\hat{W}\times D}$ is the fused patch tokens of PTF.

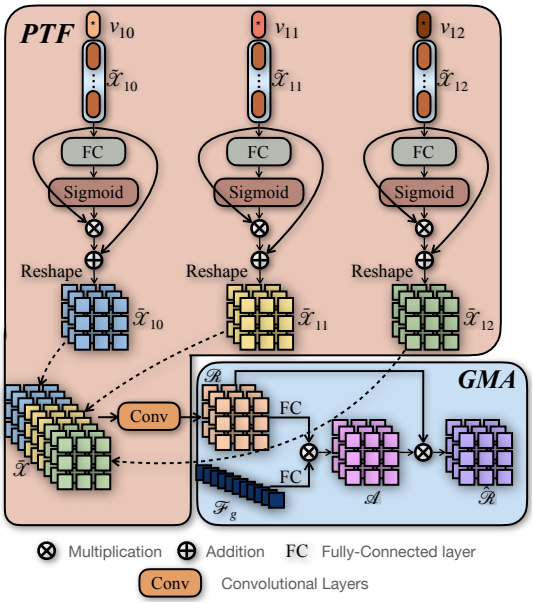

Fig. 4. Our proposed Patch Token Fusion (PTF) and Global-guided Multi-head Attention (GMA).

3.3.2. Global-guided Multi-head Attention

Although the global feature $\mathcal{F}_g$ may neglect some fine-grained information, it contains rich semantic information (Dosovitskiy et al., 2021). Considering this fact, we design a guidance mechanism to further boost the discriminative representation of the enhanced patch tokens. Specifically, as shown in the right-bottom of Fig 4, we treat the global feature $\mathcal{F}_g \in \mathbb{R}^{O\times1\times D'}$ as query. Meanwhile, the enhanced patch tokens $\mathcal{R}$ reshaped to $\mathcal{R}^{(r)} \in \mathbb{R}^{O\times N\times D'}$ are treated as key. Here, O is the number of heads and $N = O\times D'$. We set the number of heads O to be 12. Formally, the Global-guided Multi-head Attention (GMA) can be expressed as:

$$\mathcal{A}_p = Sigmoid\left((\mathcal{F}_g\times\mathcal{W}_q)\otimes(\mathcal{R}_p^{(r)}\times\mathcal{W}_k)\right), \tag{8}$$

where $\mathcal{A}_p \in \mathbb{R}^{O\times1\times D'}$ is the score of p-th patch token $\mathcal{R}_p$. $\mathcal{W}_k \in \mathbb{R}^{O\times D'\times\frac{D'}{r_2}}$ and $\mathcal{W}_q \in \mathbb{R}^{O\times D'\times\frac{D'}{r_2}}$ are learnable parameters. r_2 is a reduction ratio. Then, we can obtain the attention map $\mathcal{A} \in \mathbb{R}^{O\times N\times D'}$ by concatenating the attention scores,

$$\mathcal{A} = [\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_N]. \tag{9}$$

Finally, the resulted patch tokens $\hat{\mathcal{R}}$ are defined as:

$$\hat{\mathcal{R}} = \mathcal{R} + \mathcal{R}\times\mathcal{A}. \tag{10}$$

From the above procedure (especially Eq. 8), one can see that the patch tokens are further enhanced by the the global feature $\mathcal{F}_g$. As a result, they can deliver more discriminative features.

3.3.3. Part-based Transformer Layers

Recently, some part-based Transformer methods (Zhu et al., 2021; Zhang et al., 2021b; Yan et al., 2023) attempt to extract the local fine-grained information by dividing the feature maps and directly supervising these partition features. However, they may neglect two issues: 1) The patch tokens in Transformers contain rich semantics for global modeling. Thus, it may introduce irrelevant information by directly utilizing the partition features for training and testing. 2) The interaction between the local-wise patch tokens and global features can enhance the robustness of local representations. However, previous works neglect to incorporate global information to obtain comprehensive representations. For example, PTCR (Li et al., 2022b) utilizes a token perception module to extract local discriminative cues and leverages the powerful PVTv2 (Wang et al., 2022c) backbone to achieve impressive performance. Based on these facts, we present Part-based Transformer Layers (PTL) to extract discriminative local features. Our method goes beyond solely extracting information from the outputs of the last few Transformer layers. We incorporate the global information during the extraction of local representations, which provides additional references. The proposed method can capture both local and global cues, leading to more comprehensive and informative representations for object Re-ID.

Specifically, as shown in the bottom of Fig. 3, we split the enhanced patch tokens $\hat{\mathcal{R}}$ into three horizontal stripes and construct three local sequences. In addition, we prefix the last three class tokens v_{10}, v_{11}, v_{12} to each part sequence for further utilizing the global information. To improve the feature representation ability, we introduce an additional class token $e_t (t = 1, 2, 3)$ to each local sequence. The input for part-based transformers is formulated as:

$$\check{\mathcal{R}}_t = \left[e_t, v_{10}, v_{11}, v_{12}, x_t^{\frac{(t-1)\times N}{3}+1}, \cdots, x_t^{\frac{t\times N}{3}+1}\right], \tag{11}$$

where x_t^n is the n-th patch token in the t-th part. Then, each part-wise sequence $\check{\mathcal{R}}_t$ is followed by two independent Transformer layers, which contain a Multi-head Self-attention Layer (MSA), a Feed-Forward Network (FFN), Layer Normalizations (LN) (Ba et al., 2016) and residual connections,

$$\check{\mathcal{R}}_t' = \check{\mathcal{R}}_t + MSA\left(LN(\check{\mathcal{R}}_t)\right), \tag{12}$$

$$\check{\mathcal{R}}_t'' = \check{\mathcal{R}}_t' + FFN\left(LN(\check{\mathcal{R}}_t')\right). \tag{13}$$

Finally, to encourage the PTL to learn more diverse and complementary information, we supervise the concatenated class tokens $\mathcal{F}_l = [e_1, e_2, e_3]$ with loss functions.

With the above modules, we not only obtain local representations by dividing the feature map into local parts, but also introduce global information through multiple global class tokens (v_{10}, v_{11}, v_{12}) into the part-based self-attention mechanism.

Thus, both global and local cues are aggregated into class tokens, resulting in comprehensive feature representations.

3.4. Loss Functions

To train our proposed framework, we follow previous works and adopt the cross-entropy loss and triplet loss (Hermans et al., 2017). Specifically, the cross-entropy loss is defined as

$$\mathcal{L}_c = -\frac{1}{\mathcal{P}\times\mathcal{K}}\sum_{i=1}^{\mathcal{P}}\sum_{j=1}^{\mathcal{K}} log\frac{exp(W_{y_{i,j}}^T\cdot f_{i,j})}{\sum_{c=1}^{\mathcal{P}\cdot\mathcal{K}} exp(W_c^T\cdot f_{i,j})}, \tag{14}$$

where $\mathcal{P}$ and $\mathcal{K}$ are the number of identities and sampled images of each identity. $W_{y_{i,j}}^T$ means the weight parameters of the *i-th* label in the classification layer. $f_{i,j}$ represent the feature corresponding to the *j-th* sample of the label *i*. In addition, the triplet loss is formulated as,

$$\mathcal{L}_t = log\left[1 + exp(\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2)\right], \tag{15}$$

where f_a, f_p, and f_n are the anchor feature, positive features and negative features, respectively. Finally, the global feature $\mathcal{F}_g$, the local feature $\mathcal{F}_l$ and the class token embbeding $\mathcal{F}_{cls}$ are trained with L_c and L_t, while the class token v_{10} and v_{11} trained with L_c. The overall loss is computed as follow:

$$\mathcal{L}_{total} = \frac{1}{3}\sum_{u\in\{\mathcal{F}_g,\mathcal{F}_l,\mathcal{F}_{cls}\}}(\mathcal{L}_c(u) + \mathcal{L}_t(u)) + \frac{1}{2}\sum_{z\in\{v_{10},v_{11}\}}(\mathcal{L}_c(z)). \tag{16}$$

During testing, we concatenate global features, local features, and the last class token embbeding as the image representation,

$$\mathcal{F} = [\mathcal{F}_g, \mathcal{F}_l, \mathcal{F}_{cls}], \tag{17}$$

where $\mathcal{F}_{cls}$ is the last class token embbeding v_{12}.

4. Experiments

4.1. Datasets and Evaluation Metrics

To fully verify the effectiveness of our proposed framework, we perform experiments on four large-scale object Re-ID datasets, *i.e.*, Market1501 (Zheng et al., 2015), DukeMTMC-ReID (Ristani et al., 2016), MSMT17 (Wei et al., 2018) and VeRi-776 (Liu et al., 2016). The details of these datasets are summarized in Tab. 2. Following previous Re-ID works, we adopt mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank1 as our evaluation metrics.

4.2. Implementation Details

We implement our framework based on the PyTorch toolbox. Experimental devices include an Intel(R) Xeon(R) Platinum 8350C CPU and one NVIDIA GTX 3090 GPU (24G memory). For model training, we uniformly resize all person images to 256×128 and all vehicle images to 256×256, then followed by random cropping, horizontal flipping and random erasing (Zhong et al., 2020) as data augmentations. In addition, there are $\mathcal{B} = \mathcal{P} \times \mathcal{K}$ images sampled to the triplet loss and cross-entropy loss in a mini-batch for every training iteration. We randomly select $\mathcal{P} = 16$ identities and $\mathcal{K} = 4$ images for each identity. We employ SGD as our optimizer for total 200 epochs with a momentum of 0.9 and the weight decay of $1 \times e^{-4}$. The initial learning rate is 8×10^{-3} with a cosine learning rate decay. We adopt the ViT-B/16 (Dosovitskiy et al., 2021) with a stride 12 as our backbone, which is pre-trained on ImageNet-21k (Deng et al., 2009) and then fine-tuned on ImageNet-1k (Deng et al., 2009). We will release the source code for reproduction.

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

In Tab. 1 and Tab. 3, our GLTrans is compared with other state-of-the-art methods on four benchmarks.

Market1501: Tab. 1 shows the performances of compared methods on Market1501. Comparing with other models, we can see that our framework achieves very competitive results, especially in mAP. While the Rank1 score of our model is inferior to some compared methods, (*e.g.*, ISP (Zhu et al., 2020), HAT (Zhang et al., 2021b) and SCSN (Chen et al., 2020b)).

MSMT17: As shown in Tab. 1, our model achieves the best performance in terms of mAP and Rank1 on MSMT17. It is worth pointing out that the mAP score of our model is higher than ISP (Zhu et al., 2020) and TransReID (He et al., 2021) by 1.4% and 1.1%, respectively. The results indicate that the fusion of multi-layer features guided by global cues can obtain complementary and fine-grained feature representations.

DukeMTMC-ReID: Tab. 1 shows that our method achieves very comparable performances in terms of mAP and Rank1 on DukeMTMC-ReID. More specifically, our GLTrans surpasses TransReID (He et al., 2021), AAformer (Zhu et al., 2021) and PFD (Wang et al., 2022b) by 0.4%, 2.4%, and 0.2% on mAP score, respectively. It indicates that by exploring the complementary local and global information, our GLTrans can obtain more robust representations.

VeRi-776: To further verify the ability of our proposed model, we also compare GLTrans with some vehicle Re-ID models (He et al., 2021; Zhu et al., 2022). As shown in Tab. 3, our framework achieves the best performance in mAP and Rank1. Specifically, our GLTrans outperforms TransReID (He et al., 2021) by 0.9% and 0.4% in mAP/Rank1. In fact, vehicle Re-ID faces small inter-class variances and significant intra-class variances. This implies that recognizing local information has a significant impact on discriminative feature representations. The results in Tab. 3 show that, unlike other ViT-based methods, our proposed approach considers both local and global cues, resulting in the excellent performance.

4.4. Ablation Study

In this subsection, we conduct ablation experiments to verify the effect of the key components. The baseline method adopts a ViT-B/16 (Dosovitskiy et al., 2021) with overlapped patches and camera/viewpoint information embedding. All the ablation studies are conducted on the MSMT17 dataset. However, the results on other datasets show similar trends.

Comparisons with other part-based methods. We verify the effect of different part-based methods. Here, PCB* refers to the reproduction of PCB (Sun et al., 2018) based on ViT. Specifically, we reshape the output tokens of ViT into a feature map based on their spatial coordinates and extract local features

Table 1. Quantitative comparison of state-of-the-art methods on three public person Re-ID datasets.

Method	Backbone	Market1501		MSMT17		DukeMTMC	
		mAP	Rank1	mAP	Rank1	mAP	Rank1
SPReID (Kalayeh et al., 2018)	ResNet152	83.4	93.7	–	–	73.3	85.9
CASN (Zheng et al., 2019b)	ResNet50	82.8	94.4	–	–	73.7	87.7
BATNet (Fang et al., 2019)	ResNet50	84.7	95.1	56.8	79.5	77.3	87.7
MGN (Wang et al., 2018a)	ResNet50	86.9	95.7	–	–	78.4	88.7
ABDNet (Chen et al., 2019)	ResNet50	88.3	95.6	60.8	82.3	78.6	89.0
Pyramid (Zheng et al., 2019a)	ResNet101	88.2	95.7	–	–	79.0	89.0
OSNet (Zhou et al., 2019)	OSNet	84.9	94.8	52.9	78.7	73.5	88.6
SNR (Jin et al., 2020b)	ResNet50	84.7	94.4	–	–	73.0	85.9
RGA-SC (Zhang et al., 2020b)	ResNet50	88.4	**96.1**	57.5	80.3	–	–
SCSN (Chen et al., 2020b)	ResNet50	88.5	95.7	–	–	79.0	**91.0**
CDNet (Li et al., 2021a)	CDNet	86.0	95.1	54.7	78.9	76.8	88.6
PAT (Li et al., 2021c)	ResNet50	88.0	95.4	–	–	78.2	88.8
ISP (Zhu et al., 2020)	HRNet48	88.6	95.3	–	–	80.0	89.6
FED (Wang et al., 2022d)	ResNet50	86.3	95.0	–	–	78.0	89.4
Nformer (Wang et al., 2022a)	ResNet50	**91.1**	94.7	59.8	77.3	**83.5**	89.4
AAformer (Zhu et al., 2021)	ViT-B/16	87.7	95.4	62.6	83.1	80.0	90.1
TransReID (He et al., 2021)	ViT-B/16	88.9	95.2	67.4	85.3	82.0	**90.7**
APD (Lai et al., 2021b)	ResNet50	87.5	95.5	57.1	79.8	74.2	87.1
HAT (Zhang et al., 2021b)	ResNet50	89.8	95.8	61.2	82.3	81.4	90.4
ADSO (Zhang et al., 2021a)	ResNet50	87.7	94.8	–	–	74.9	87.4
PFD (Wang et al., 2022b)	ViT-B/16	89.6	95.5	65.1	82.7	82.2	90.6
DCAL (Zhu et al., 2022)	ViT-B/16	87.5	94.7	64.0	83.1	80.1	89.0
GLTrans (Ours)	ViT-B/16	90.0	95.6	**69.0**	**85.8**	82.4	**90.7**

Table 2. Statistics of used datasets.

Dataset	Object	ID	Image	Cam(View)
MSMT17	Person	4,101	126,441	15
Market1501	Person	1,501	32,668	6
DukeMTMC-ReID	Person	1,404	36,441	8
VeRi-776	Vehicle	776	49,357	20(8)

Table 3. Quantitative comparison of state-of-the-art methods on VeRi-776.

Method	Backbone	VeRi-776	
		mAP	Rank1
PPReID (He et al., 2019)	ResNet50	72.5	93.3
SAN (Qian et al., 2020)	ResNet50	72.5	93.3
UMTS (Jin et al., 2020a)	ResNet50	75.9	95.8
VANet (Chu et al., 2019)	ResNet50	66.3	95.8
SPAN (Chen et al., 2020a)	ResNet50	68.9	94.0
PGAN (Zhang et al., 2020a)	ResNet50	79.3	96.5
PVEN (Meng et al., 2020)	ResNet50	79.5	95.6
SAVER (Khorramshahi et al., 2020)	ResNet50	79.6	96.4
CFVMNet (Sun et al., 2020)	ResNet50	77.1	95.3
GLAMOR (Suprem and Pu, 2020)	ResNet50	80.3	96.5
MPC (Li et al., 2021b)	ResNet50	80.9	96.2
MsKAT (Li et al., 2022a)	ResNet50	82.0	97.1
TransReID (He et al., 2021)	ViT-B/16	82.0	97.1
DCAL (Zhu et al., 2022)	ViT-B/16	80.2	96.9
GLTrans (Ours)	ViT-B/16	**82.9**	**97.5**

by applying the average pooling to the divided feature map. During testing, we concatenate the averaged features with class tokens. MSF* indicates that we aggregate tokens from multiple layers using GAE and PTF modules to obtain global and local representations. The local representations are obtained by applying the average pooling to the fused feature map.

As shown in Tab. 4, compared with the baseline, it clearly achieves improvements by using the part-aware method and multi-layer fusion. The results indicate that the patch tokens contain valuable fine-grained information. In addition, the outputs of the last few layers of ViT complement each other. Although PCB* can explore local cues from patch tokens, it can not fully explore complementary local cues by solely using the outputs of the last layer. Fig. 2 and Fig. 5 clearly show these facts. The results shown in the last row of Tab. 4 illustrate that simultaneously utilizing the part modeling methods and multi-stage fusion mechanisms achieve best performances.

Table 4. Performance comparison with different part-based methods.

Model	mAP	Rank1
Baseline	66.3	84.0
PCB*	68.2	85.2
MSF*	67.6	84.8
Ours	**69.0**	**85.8**

The impacts of different components. In Tab. 5, we present the ablation studies of GAE and LMF modules. Model-1 corresponds to the baseline. Compared with Model-2, we can see that employing the GAE significantly improves the perfor-

mance over the baseline by 1.5% on mAP and 0.8% on Rank1 accuracy. The main reason is that the GAE can aggregate several class tokens, representing multiple global semantic information of input images. Model-3 means that we add the PTL to obtain the local information. One can see that the performance further improves by 0.7% and 0.3% in terms of mAP and Rank1, respectively. The main reason is that our PTL makes the model obtain a more robust representation by focusing on more detailed information. Model-4 and Model-5 mean that we further introduce the PTF and GMA. As shown in Tab. 5, the final model further provides 0.5% mAP and 0.3% Rank1 accuracy improvements, respectively. It indicates that aggregating and enhancing the multi-layer patch tokens can explore more discriminative and robust features. The above results clearly demonstrate the effectiveness of our proposed modules.

Table 5. Performance comparison of key components.

Model	Baseline	GAE	LMF			MSMT17	
			PTL	PTF	GMA	mAP	Rank1
1	✓	×	×	×	×	66.3	84.0
2	✓	✓	×	×	×	67.8	85.2
3	✓	✓	✓	×	×	68.5	85.5
4	✓	✓	✓	✓	×	68.8	85.7
5	✓	✓	✓	✓	✓	**69.0**	**85.8**

Table 6. Performance comparison with different aggregated layers.

Layer	mAP	Rank1
9,10,11,12	68.2	85.2
10,11,12	**68.6**	**85.8**
11, 12	67.8	85.1
12	67.5	85.1

The impacts of aggregating different layers. The choice of which layers to aggregate is an important hyper-parameter in our framework. Therefore, we investigate the impacts of aggregating different layers. As shown in Tab. 6, we find that the highest performance is achieved by aggregating the last three layers of ViT. It is verified that the features from the last few layers actually contain different semantic information. They can complement and enhance each other.

The impacts of different aggregation strategies in GAE. We exhibit experimental results of four aggregation strategies in Tab. 7. We first adopt addition and concatenation to aggregate the features, respectively. As shown in the 2-*th* and 3-*th* rows, the performances are not satisfactory. The main reason is that multiple features have different semantics. This kind of straightforward methods can not explore the information among them and even destroy the feature distributions. Meanwhile, in the 4-*th* and 5-*th* rows, we illustrate the results of employing different FC layers. As can be observed, the best performance is achieved by using only one FC layer. We argue that the FC layer can consider the relation between the multiple features with different semantics. One FC layer is more easily trained and avoids over-fitting.

The impacts of different partition numbers. It is of great interest to determine the optimal partition strategy for the feature map to extract local fine-grained features. To this end, we conduct experiments to assess the impacts of varying the number of parts. As depicted in Tab. 8, dividing the feature map into more parts does not lead to performance improvement. For instance, when the feature map is divided into six parts, the performance is comparable to that of three parts. Consequently, in this paper, to keep the computational efficiency, we divide the feature map into three parts.

Table 7. Performance comparison with different aggregation strategies.

Stage	mAP	Rank1
Addition	66.4	84.7
Concatenation	66.6	84.5
One FC	**69.0**	**85.8**
Two FC	66.3	84.1

Table 8. Performance comparison with different partition numbers.

Number	mAP	Rank1
1	68.3	85.2
2	69.0	85.7
3	**69.0**	**85.8**
6	68.9	85.7

The impacts of shared Transformers. Considering the computational efficiency, it is crucial to evaluate the configuration of Transformer blocks in LMF. As illustrated in Tab. 9, we conduct several experiments with different settings. Specifically, Model-1 indicates that we divide the feature map and utilize an average pooling to obtain local feature representations. Model-2 and Model-3 mean that we employ one and two unshared Transformer layers to extract the local feature representation, respectively. Model-4 and Model-5 means that we employ one and two shared Transformer layers to extract the local feature representation, respectively. As can be observed, Model-5 achieves the best performance. We believe that the improvements can be attributed to the fusion of multi-layer semantic features. In addition, more shared Transformers enable the extraction of diverse information from multiple regions.

Table 9. Performance comparison with different settings in LMF.

Model	Trans	Shared	Layer	mAP	Rank1
1	×	×	0	68.2	85.5
2	✓	×	1	68.9	85.7
3	✓	×	2	67.6	85.1
4	✓	✓	1	68.5	85.5
5	✓	✓	2	**69.0**	**85.8**

The impacts of different heads of GMA. To acquire more discriminative token representations, we introduce multiple heads in GMA. To verify the impacts, we conduct several experiments, as presented in Tab. 10. As can be observed, as the number of heads increases, the performance improves. The results indicate that the multi-head attention mechanism can extract multiple semantic information and enhance the guided effect of global information.

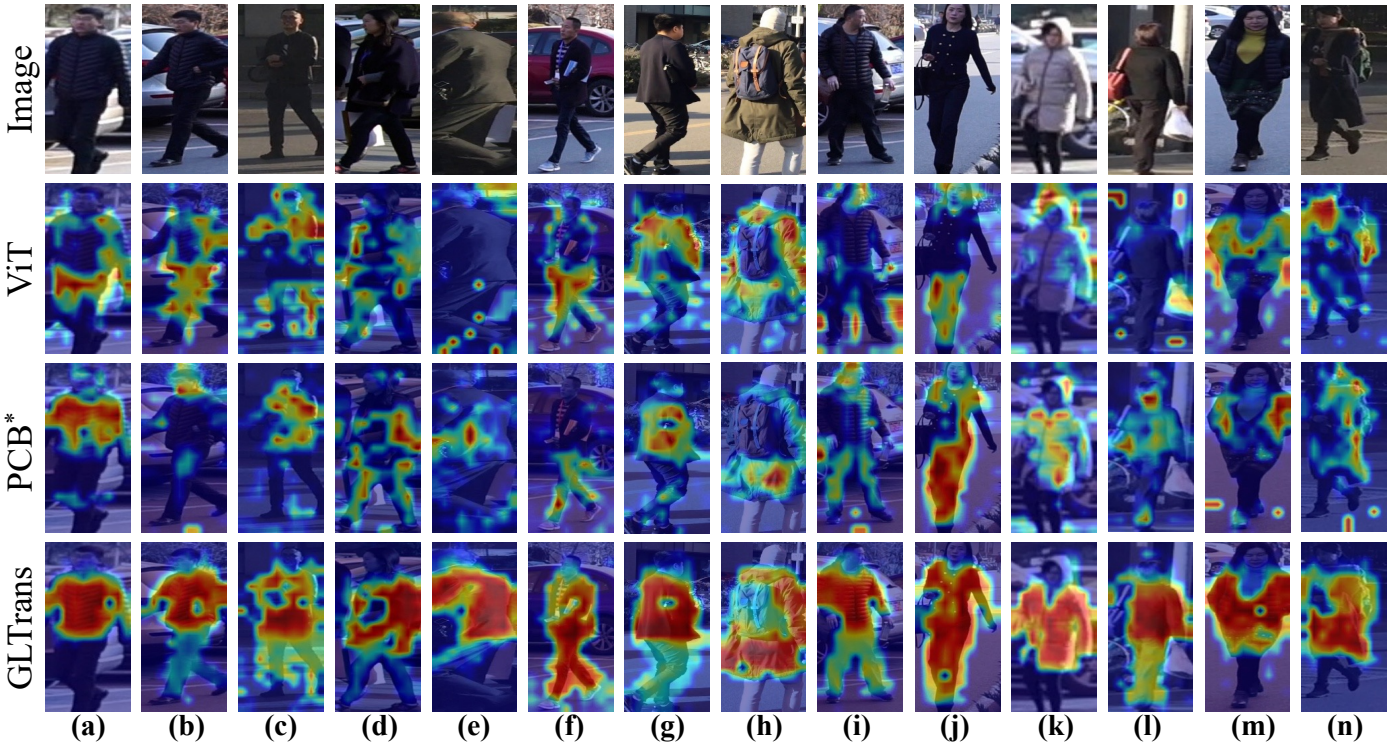

Fig. 5. Visualization of the differences between ViT, PCB* and GLTrans by Grad-CAM (Selvaraju et al., 2017). Deeper red colors signify higher weights. The first row is the input images. The second, third and fourth rows are the activation maps produced by ViT, PCB* and our GLTrans, respectively.

Table 10. Performance comparison with different heads of GMA.

Number	mAP	Rank1
1	68.5	85.4
4	68.6	85.4
6	68.8	85.3
12	**69.0**	**85.8**

4.5. Qualitative Results

4.5.1. Visualization of Activation Maps

To qualitatively analyze our model, we present the visual results of ViT (Dosovitskiy et al., 2021), PCB* and our GLTrans on the MSMT17 dataset. As illustrated in the second row of Fig. 5, the scattered salient regions suggest that the ViT usually focuses on backgrounds and unrelated objects. In contrast, the incorporation of part-based mechanisms can encourage the model to emphasize regions more pertinent to the target individual. For instance, as depicted in Fig. 5(e), (i), (k) and (l), the PCB* model discerns more semantically meaningful cues. However, Fig. 5(c), (f) and (h) demonstrate that PCB* relies on a few salient regions and miss useful cues due to the absence of global guidance. Furthermore, our GLTrans demonstrates the effect of global guidance, allowing the model to focus on more discriminative and holistic information related to the person. For example, as illustrated in Fig. 5(f) and (i), our GLTrans not only identifies meaningful local regions but also mitigates the influence of extraneous information.

4.5.2. Visualization of Retrieval Results

To further demonstrate the superiority of our proposed model, we present the retrieval results obtained by the ViT and our GLTrans. As shown in Fig. 6, our GLTrans can get more hard positive samples and discard hard negative ones even if the global appearances of persons are similar. For example, in the 3-*th* row, the baseline model only concerns the person with a blue coat and black pants. It misses the local cues (*e.g.* gender and handbag). As shown in the 4-*th* row, our GLTrans obtains more robust and complementary representations, by refining both the global and local features.

Meanwhile, as shown in the 5-*th* row, the baseline model can only retrieve the samples with the black coat and pants. It misses local cues about the orange knapsack. The main reason is that the baseline focuses on the global-view information and ignores the discriminative local features. In contrast, our GLTrans achieves better results. It considers both the global information from class tokens and the local details from patch tokens. The same phenomenon can be found in other examples. These visual results clearly illustrate that our GLTrans exploits both local and global information more effectively.

5. Conclusion

In this paper, we propose a novel learning framework named GLTrans for image-based object Re-ID. To obtain complementary global representations, we propose a Global Aggregation Encoder (GAE) to aggregate multi-layer class tokens of ViT. To mine discriminative local cues, we introduce a Local Multi-layer Fusion (LMF). It contains three main modules, *i.e.*,

Fig. 6. Retrieval results with three query samples on MSMT17. For each query, the first and second row are the ranking list produced by ViT and our GLTrans, respectively. The black, green and red boxes mean the query sample, true positive and false positive, respectively.

Patch Tokens Fusion (PTF), Global-guided Multi-head Attention (GMA) and Part-based Transformer Layers (PTL). The PTF can adaptively enhance and fuse multi-layer patch tokens. The GMA enhance the local patch tokens guided by the global-view. The PTL adopts a local Transformer, encouraging more diverse and complementary local information. The proposed framework takes the global and local advantages of vision Transformers for robust object Re-ID. Experiments on four large-scale object Re-ID benchmarks demonstrate that our method achieves better performance than most state-of-the-art methods. In the future, we will reduce the computation and improve the representation ability of diverse Transformers.

References

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv:1607.06450 .

Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., Lin, C.W., 2022. Structure-aware positional transformer for visible-infrared person re-identification. TIP 31, 2352–2364.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z., 2019. Abd-net: Attentive but diverse person re-identification, in: CVPR.

Chen, T.S., Liu, C.T., Wu, C.W., Chien, S.Y., 2020a. Orientation-aware vehicle re-identification with semantics-guided part attention network, in: ECCV.

Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R., Yang, Y., 2020b. Salience-guided cascaded suppression network for person re-identification, in: CVPR.

Chen, X., Xu, J., Xu, J., Gao, S., 2021. Oh-former: Omni-relational high-order transformer for person re-identification. arXiv preprint arXiv:2109.11159 .

Chu, R., Sun, Y., Li, Y., Liu, Z., Zhang, C., Wei, Y., 2019. Vehicle re-identification with viewpoint-aware metric learning, in: CVPR.

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport, in: NIPS.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image isworth 16x16 words: Transformers for image recognition at scale., in: ICLR.

Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M., 2019. Bilinear attention networks for person retrieval, in: CVPR.

Floridi, L., Chiriatti, M., 2020. Gpt-3: Its nature, scope, limits, and consequences, in: MM.

Gao, S., Yu, C., Zhang, P., Lu, H., 2024. Part representation learning with teacher-student decoder for occluded person re-identification, in: ICASSP, pp. 2660–2664.

He, B., Li, J., Zhao, Y., Tian, Y., 2019. Part-regularized near-duplicate vehicle re-identification, in: CVPR.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR.

He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W., 2021. Transreid: Transformer-based object re-identification, in: ICCV.

Hendrycks, D., Gimpel, K., 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units, in: ICLR.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv:1703.07737 .

Huang, M., Hou, C., Yang, Q., Wang, Z., 2023. Reasoning and tuning: Graph attention network for occluded person re-identification. TIP 32, 1568–1582.

Jin, X., Lan, C., Zeng, W., Chen, Z., 2020a. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification, in: AAAI.

Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L., 2020b. Style normalization and restitution for generalizable person re-identification, in: CVPR.

Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M., 2018. Human semantic parsing for person re-identification, in: CVPR.

Khatun, A., Denman, S., Sridharan, S., Fookes, C., 2020. Joint identification–verification for person re-identification: A four stream deep learning approach with improved quartet loss function, in: CVIU.

Khorramshahi, P., Peri, N., Chen, J.c., Chellappa, R., 2020. The devil is in the details: Self-supervised attention for vehicle re-identification, in: ECCV.

Lai, S., Chai, Z., Wei, X., 2021a. Transformer meets part model: Adaptive part division for person re-identification, in: ICCV.

Lai, S., Chai, Z., Wei, X., 2021b. Transformer meets part model: Adaptive part division for person re-identification, in: CVPR.

Li, H., Li, C., Zheng, A., Tang, J., Luo, B., 2022a. Mskat: Multi-scale knowledge-aware transformer for vehicle re-identification, in: TITS.

Li, H., Wu, G., Zheng, W.S., 2021a. Combined depth space based architecture search for person re-identification, in: CVPR.

Li, H., Ye, M., Wang, C., Du, B., 2022b. Pyramidal transformer with conv-patchify for person re-identification, in: ACM MM, pp. 7317–7326.

Li, M., Liu, J., Zheng, C., Huang, X., Zhang, Z., 2021b. Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification, in: TMM.

Li, W., Zhu, X., Gong, S., 2018. Harmonious attention network for person re-identification, in: CVPR.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F., 2021c. Diverse part discovery: Occluded person re-identification with part-aware transformer, in: CVPR.

Lin, X., Zhu, L., Yang, S., Wang, Y., 2023. Diff attention: A novel attention scheme for person re-identification, in: CVIU.

Liu, X., Liu, W., Ma, H., Fu, H., 2016. Large-scale vehicle re-identification in urban surveillance videos, in: ICME.

Liu, X., Yu, C., Zhang, P., Lu, H., 2023a. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. TNNLS .

Liu, X., Zhang, P., Yu, C., Lu, H., Qian, X., Yang, X., 2021. A video is worth three views: Trigeminal transformers for video-based person re-identification. arXiv preprint arXiv:2104.01745 .

Liu, Z., Mu, X., Lu, Y., Zhang, T., Tian, Y., 2023b. Learning transformer-based attention region with multiple scales for occluded person re-identification, in: CVIU.

Lu, H., Zou, X., Zhang, P., 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification, in: AAAI, pp. 1835–1843.

Meng, D., Li, L., Liu, X., Li, Y., Yang, S., Zha, Z.J., Gao, X., Wang, S., Huang, Q., 2020. Parsing-based view-aware embedding network for vehicle re-identification, in: CVPR.

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q., 2021. Conformer: Local features coupling global representations for visual recognition, in: CVPR.

Qian, J., Jiang, W., Luo, H., Yu, H., 2020. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification, in: MST.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: ECCV.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: ICCV.

Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G., 2018. Dual attention matching network for context-aware feature sequence based person re-identification, in: CVPR.

Song, C.H., Yoon, J., Choi, S., Avrithis, Y., 2023. Boosting vision transformers for image retrieval, in: WACV.

Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M., 2018. Part-aligned bilinear representations for person re-identification, in: ECCV.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019a. Deep high-resolution representation learning for human pose estimation, in: CVPR.

Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J., 2019b. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, in: CVPR.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: ECCV.

Sun, Z., Nie, X., Xi, X., Yin, Y., 2020. Cfvmnet: A multi-branch network for vehicle re-identification based on common field of view, in: ACM MM.

Suprem, A., Pu, C., 2020. Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention. arXiv preprint arXiv:2002.02256 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: NIPS.

Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X., 2018a. Learning discriminative features with multiple granularities for person re-identification, in: ACM MM.

Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E., 2022a. Nformer: Robust person re-identification with neighbor transformer, in: CVPR.

Wang, T., Liu, H., Song, P., Guo, T., Shi, W., 2022b. Pose-guided feature disentangling for occluded person re-identification based on transformer, in: AAAI.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022c. Pvt v2: Improved baselines with pyramid vision transformer. CVM 8, 415–424.

Wang, X., Girshick, R., Gupta, A., He, K., 2018b. Non-local neural networks, in: CVPR.

Wang, Y., Chen, Z., Wu, F., Wang, G., 2018c. Person re-identification with cascaded pairwise convolutions, in: CVPR.

Wang, Y., Liu, X., Zhang, P., Lu, H., Tu, Z., Lu, H., 2024. Top-reid: Multi-spectral object re-identification with token permutation, in: AAAI, pp. 5758–5766.

Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J., 2022d. Feature erasing and diffusion network for occluded person re-identification, in: CVPR.

Wei, L., Zhang, S., Gao, W., Tian, Q., 2018. Person transfer gan to bridge domain gap for person re-identification, in: ICCV.

Weng, J., Hu, K., Yao, T., Wang, J., Wang, Z., 2023. Federated unsupervised cluster-contrastive learning for person re-identification: A coarse-to-fine approach, in: CVIU.

Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W., 2018. Attention-aware compositional network for person re-identification, in: CVPR.

Yan, P., Liu, X., Zhang, P., Lu, H., 2023. Learning convolutional multi-level transformers for image-based person re-identification. Visual Intelligence 1, 24.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2021. Deep learning for person re-identification: A survey and outlook, in: TPAMI.

Ye, M., Wu, Z., Chen, C., Du, B., 2024. Channel augmentation for visible-infrared re-identification. TPAMI 46, 2299–2315.

Yu, C., Liu, X., Wang, Y., Zhang, P., Lu, H., 2024. Tf-clip: Learning text-free clip for video-based person re-identification, in: AAAI, pp. 6764–6772.

Zang, X., Li, G., Gao, W., 2022. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval, in: TII.

Zhang, A., Gao, Y., Niu, Y., Liu, W., Zhou, Y., 2021a. Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck, in: CVPR.

Zhang, G., Zhang, P., Qi, J., Lu, H., 2021b. Hat: Hierarchical aggregation transformers for person re-identification, in: ACM MM.

Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H., 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification. arXiv preprint arXiv:2403.10254 .

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J., 2017. Alignedreid: Surpassing human-level performance in person re-identification. arXiv:1711.08184 .

Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., Shen, C., 2020a. Part-guided attention learning for vehicle instance retrieval, in: TITS.

Zhang, Y., Kang, W., Liu, Y., Zhu, P., 2023. Multi-scale semantic and detail extraction network for lightweight person re-identification. CVIU 236, 103813.

Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z., 2020b. Relation-aware global attention for person re-identification, in: CVPR.

Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R., 2019a. Pyramidal person re-identification via multi-loss dynamic training, in: CVPR.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark, in: ICCV.

Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: Past, present and future. arXiv:1610.02984 .

Zheng, M., Karanam, S., Wu, Z., Radke, R.J., 2019b. Re-identification with consistent attentive siamese networks, in: CVPR.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation, in: AAAI.

Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019. Omni-scale feature learning for person re-identification, in: CVPR.

Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., Shan, Y., 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification, in: CVPR.

Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J., 2020. Identity-guided human semantic parsing for person re-identification, in: ECCV.

Zhu, K., Guo, H., Zhang, S., Wang, Y., Huang, G., Qiao, H., Liu, J., Wang, J., Tang, M., 2021. Aaformer: Auto-aligned transformer for person re-identification. arXiv:2104.00921 .