Transformer Based Multi-Grained Features
for Unsupervised Person Re-Identification

Jiachen Li[1], Menglin Wang[2], and Xiaojin Gong[*]

[1,2,*]College of Information Science and Electronic Engineering, Zhejiang University, China

[1,*]{lijiachen_isee, gongxj}@zju.edu.cn, [2]lynnwang6875@gmail.com;

Abstract

Multi-grained features extracted from convolutional neural networks (CNNs) have demonstrated their strong discrimination ability in supervised person re-identification (Re-ID) tasks. Inspired by them, this work investigates the way of extracting multi-grained features from a pure transformer network to address the unsupervised Re-ID problem that is label-free but much more challenging. To this end, we build a dual-branch network architecture based upon a modified Vision Transformer (ViT). The local tokens output in each branch are reshaped and then uniformly partitioned into multiple stripes to generate part-level features, while the global tokens of two branches are averaged to produce a global feature. Further, based upon offline-online associated camera-aware proxies (O2CAP) that is a top-performing unsupervised Re-ID method, we define offline and online contrastive learning losses with respect to both global and part-level features to conduct unsupervised learning. Extensive experiments on three person Re-ID datasets show that the proposed method outperforms state-of-the-art unsupervised methods by a considerable margin, greatly mitigating the gap to supervised counterparts. Code will be available soon at `https://github.com/RikoLi/WACV23-workshop-TMGF`.

1. Introduction

Purely unsupervised person re-identification (Re-ID) aims to learn a Re-ID model without using any identity labels. This task has attracted extensive research interest because of its label-free manner, which makes it more practical and scalable to real-world deployments. Over the past few years, significant progress has been made, mainly due to the leverage of pseudo labeling [24] and contrastive learning [5, 15, 44] techniques. Existing unsupervised methods often focus on the design of various contrastive losses [3, 8, 38, 39] and the refinement of noisy pseudo labels [7, 39, 43, 51]. Most of them pay little attention to the improvement of their feature extraction networks, which are crucial for identification as well.

On the contrary, the architecture of feature extraction backbones has been extensively investigated in supervised person Re-ID. For example, besides bag of tricks (BoT) [27], partition-based [6, 35] or multi-granularity [37, 52] networks are developed to capture fine-grained cues, and attention schemes [4, 22, 33] are integrated to concentrate on discriminative parts. Recently, self-attention or transformer mechanisms [17,21,23,28,32,50,58] have also been successfully applied to supervised Re-ID. Some of them [21, 23, 50] integrate transformers with convolutional neural networks (CNNs) and the others [17,28,32,58] construct pure transformer architectures to explore long-range contexts. It has been validated that both fine-grained cues and long-range contexts greatly boost the performance of supervised Re-ID.

Inspired by the techniques developed in supervised Re-ID, especially by the CNN-based Multiple Granularity Network (MGN) [37] and the pure transformer networks [17, 28, 32, 58], we intend to investigate the way of extracting multi-grained features from a pure transformer network to address the more challenging unsupervised Re-ID problem. To this end, we take the TransReID-SSL [17, 28] network as our backbone, which builds upon the Vision Transformer (ViT) [9] but slightly modifies ViT to adapt to the Re-ID task. Then, we construct a dual-branch architecture based on the transformer backbone. Each branch duplicates the last transformer layer to produce outputs independently. The output local tokens in each branch, which are corresponding to input patches [2, 32], are reshaped and uniformly partitioned into a number of stripes to learn part-level features. Meanwhile, the output global tokens of both branches are averaged to generate a global feature. By this

---

*The corresponding author.

means, multi-grained features are effectively learned from the pure transformer network.

In order to take advantage of the learned multi-grained features for unsupervised Re-ID, we choose offline-online associated camera-aware proxies (O2CAP) [39] as our learning framework. O2CAP [39] is a state-of-the-art unsupervised Re-ID method based on a CNN backbone. We extend it by replacing its backbone with our transformer-based dual-branch network for feature extraction. In addition, we define offline and online contrastive learning losses with respect to both global and part-level features.

Although some recent works also attempt to learn part-level features from pure transformer networks [32, 58] for supervised Re-ID or leverage CNN-based part-level features for unsupervised Re-ID [7, 13, 25, 47, 48], our method distinguishes itself from them in the following aspects:

- Inspired by MGN [37], we design a dual-branch architecture appended to a pure transformer network to learn features at multiple granularities, which is simple but effective to mine fine-grained cues and capture long-range contexts at the same time.

- Based upon O2CAP [39], we additionally define part-feature based offline and online contrastive learning losses to leverage the learned multi-grained features for unsupervised Re-ID, boosting the performance significantly.

- Extensive experiments on three person Re-ID datasets show that our method outperforms state-of-the-art unsupervised methods by a considerable margin, greatly mitigating the gap to supervised counterparts.

2. Related Work

2.1. Unsupervised Person Re-ID

Previous unsupervised Re-ID methods can be roughly grouped into unsupervised domain adaptation (UDA)-based [11, 14, 18, 42, 53] or purely unsupervised [3, 8, 24, 38, 39, 46, 49] categories. In recent years, the purely unsupervised Re-ID has attracted more research interest due to its promising performance and no use of extra labeled datasets. Most unsupervised methods are clustering-based, taking clustering to produce pseudo labels and training a Re-ID model under the supervision of pseudo labels iteratively. Advanced performance has been recently achieved via the design of clustering techniques [49, 55], the leverage of contrastive learning techniques [8, 38, 39], the refinement of noisy pseudo labels [7, 43, 51], etc. Although numerous methods have been developed, almost all of them leave their CNN-based feature extraction backbones unchanged. An exception is TransReID-SSL [28], which attempts to exploit a pure transformer network for unsupervised Re-ID and significantly boosts the performance.

2.2. Part-based Person Re-ID

Part-based features can encode fine-grained cues that are shown to be helpful for discriminating IDs under the supervised setting. Various methods, which adopt direct partition [6, 35], multiple granularities [37, 52], attention schemes [22, 33], or transformer-based techniques [21, 23, 32, 50, 58], have been developed to boost the performance of supervised Re-ID. In contrast, there are only a few studies on exploiting part-level features for unsupervised person Re-ID. These unsupervised methods either design part-feature based losses [13, 47, 48], or integrate part-level features for similarity measurement [25], or use part features to refine pseudo labels [7]. In these unsupervised methods, part-level features are all extracted from CNN backbones.

2.3. Transformer-based Person Re-ID

Transformer has demonstrated its great potential in various vision tasks [9, 26]. It has also been applied to person Re-ID in recent years, mostly under the supervised setting. Initial methods [21, 23, 50] often integrate transformer layers with CNN backbones to capture fine-grained cues and long-range contexts. For instance, TPM [21] constructs a transformer-based module to adaptively merge parts generated from CNN-based networks such as PCB [35] or MGN [37]. HAT [50] designs a hierarchical aggregation transformer upon ResNet-50 [16] to integrate low-level details with high-level semantics. PAT [23] appends a part-aware transformer after a CNN backbone to discover diverse parts. Recently, pure transformer architectures, such as TransReID [17], AAformer [58], and LA-Transformer [32] are also developed for supervised Re-ID. AAformer [58] introduces part tokens to learn part features in transformer while LA-Transformer [32] designs a PCB-like strategy to extract part-level features. In contrast to them [32, 58], we construct a dual-branch architecture appended to a pure transformer to learn features at multiple granualities and utilize the multi-grained features for unsupervised Re-ID.

3. The Proposed Method

This work aims to investigate the way of extracting multi-grained features from a pure transformer network to address unsupervised Re-ID. To this end, we take a modified Vision Transformer (ViT) [9, 17, 28] as our backbone and construct a dual-branch architecture appended to the backbone for feature extraction. The last transformer layer is duplicated for each branch to produce outputs independently. Then, the local tokens in each branch are reshaped and uniformly partitioned into multiple stripes to yield part-level features, and meanwhile the global tokens of both branches are averaged to get a global feature. Further, based upon O2CAP [39], offline and online contrastive learning

losses with respect to both global and part-level features are defined for unsupervised learning. During test time, only the global feature obtained from two global tokens is used for inference. Figure 1 illustrates the entire framework.

3.1. Transformer Backbone

We choose the pure transformer architecture used in TransReID-SSL [28] as our backbone. It builds upon Vision Transformer (ViT) [9] but makes two modifications to adapt to person Re-ID. Specifically, as shown in Figure 2, an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ is first passed through an Instance-Batch Normalization (IBN)-based convolution stem to produce a feature map $\mathbf{x}' \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, in which H and W are the height and width of the image and C is the number of channels. The IBN-based convolution stem, inspired by IBN-Net [29] that is extensively used in CNN-based Re-ID methods, takes place of the vanilla convolution stem [45] which is previously added in ViT to further improve training stability and generalization ability. Then, the feature map is split into $N = \frac{HW}{P^2}$ non-overlapping patches and each patch is in size of $\frac{P}{2} \times \frac{P}{2}$. Further, each patch is projected into a D-dimensional feature $\mathbf{f} \in \mathbb{R}^D$ as an embedded token. A learnable class token cls is prepended to the sequence of patch tokens. Finally, position embeddings, together with additional camera embeddings [17, 28], are added to the embedded patch tokens to form the input for a transformer network. The entire procedure of tokenization is formally defined as follows:

$$
\begin{aligned}
\mathbf{f}_i &= \psi_i \left(\text{ICS} \left(\mathbf{x} \right) \right), \quad i = 1, ..., N; \\
\mathbf{z}^0 &= [cls; \, \mathbf{f}_1; \, ...; \, \mathbf{f}_N] + \mathbf{p} + \lambda_c \mathbf{c}.
\end{aligned}
\tag{1}
$$

Here, ICS denotes the IBN-based convolution stem and ψ_i is the partition and projection operation. $\mathbf{p} \in \mathbb{R}^{(N+1) \times D}$ is the position embedding, $\mathbf{c} \in \mathbb{R}^{(N+1) \times D}$ is the camera embedding, and λ_c is a hyper-parameter for weighting.

The embedded tokens $\mathbf{z}^0$ are then input to the transformer network composed of L transformer layers exactly as ViT [9]. Each layer consists of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) modules. The computation of layer $l \in \{1, ..., L\}$ is defined as follows:

$$
\begin{aligned}
\hat{\mathbf{z}}^{l-1} &= \text{MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \\
\mathbf{z}^l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^{l-1})) + \hat{\mathbf{z}}^{l-1},
\end{aligned}
\tag{2}
$$

in which LN denotes layer normalization. Therefore, we denote the final output of the transformer network as

$$
\mathbf{z}^L = \left[cls^L; \, \mathbf{f}_1^L; \, ...; \, \mathbf{f}_N^L \right].
\tag{3}
$$

3.2. Multiple Granularity Architecture

Inspired by the CNN-based Multiple Granularity Network (MGN) [37], we construct a dual-branch architecture to extract features at multiple granularities. The architecture is illustrated in Figure 1. We modify the above-introduced transformer backbone by duplicating the L-th transformer layer while keeping all previous layers unchanged. Each branch has one copy of the L-th transformer layer, which outputs one global token cls^L together with N local tokens $\left[\mathbf{f}_1^L; \, ...; \, \mathbf{f}_N^L \right]$. As pointed out in [2, 32], the local tokens substantially correspond to the original input patches and hence encode fine-grained local information. We therefore reshape the local tokens in each branch and partition them into horizontal stripes to generate part-level features.

More specifically, we denote the output local tokens in the i-th branch as $\left[\mathbf{f}_{i,1}^L; \, ...; \, \mathbf{f}_{i,N}^L \right]$, which are reshaped into a feature map $\mathbf{f}_i'$ in size of $\frac{H}{P} \times \frac{W}{P} \times D$. Then, the reshaped feature map is uniformly split into K_i horizontal parts and each part gets a D-dimensional feature by average pooling all features belonging to this part. Formally, this procedure is described as

$$
\begin{aligned}
\mathbf{f}_i' &= \text{Reshape} \left(\left[\mathbf{f}_{i,1}^L; \, ...; \, \mathbf{f}_{i,N}^L \right] \right), \quad i = 1, 2; \\
\mathbf{h}_{i,k} &= \text{AvgPool} \left(\text{Split} \left(\mathbf{f}_i', k \right) \right), \quad k = 1, ..., K_i.
\end{aligned}
\tag{4}
$$

Here, $\mathbf{h}_{i,k}$ denotes the k-th part-level feature in the i-th branch. With a bit change of the subscripts, we denote the set of all part features as $\{\mathbf{h}^k\}_{k=1}^K$, where $K = K_1 + K_2$.

Note that each branch also outputs a global token. Therefore, different from MGN [37] that additionally constructs a global branch, we get a global feature by simply averaging the global tokens produced in both branches. That is,

$$
\mathbf{g} = \frac{1}{2} \left(cls_1^L + cls_2^L \right).
\tag{5}
$$

Both the global feature and part-level features are further passed through a batch normalization (BN) layer and a L_2 normalization layer to have unit norms.

3.3. Unsupervised Re-ID Losses

We intend to leverage multi-grained features to promote the performance of unsupervised Re-ID. In this work, we build our losses upon O2CAP [39], which is a state-of-the-art unsupervised method using CNN-extracted global features. We extend it to include both global and part-level features for learning.

Given an unlabeled dataset, O2CAP performs the unsupervised learning by conducting a clustering step and a model learning step alternatively and iteratively. In each clustering step, it utilizes DBSCAN [10] to cluster all images with respect to their global features, and then split each cluster into multiple camera-aware proxies according to camera information. The proxies are taken as pseudo labels for supervision. Let us denote the dataset with pseudo labels as $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^{N_I}$, where image $\mathbf{x}_i$ extracts a global feature $\mathbf{g}_i$ and a set of part-level features $\{\mathbf{h}_i^k\}_{k=1}^K$.
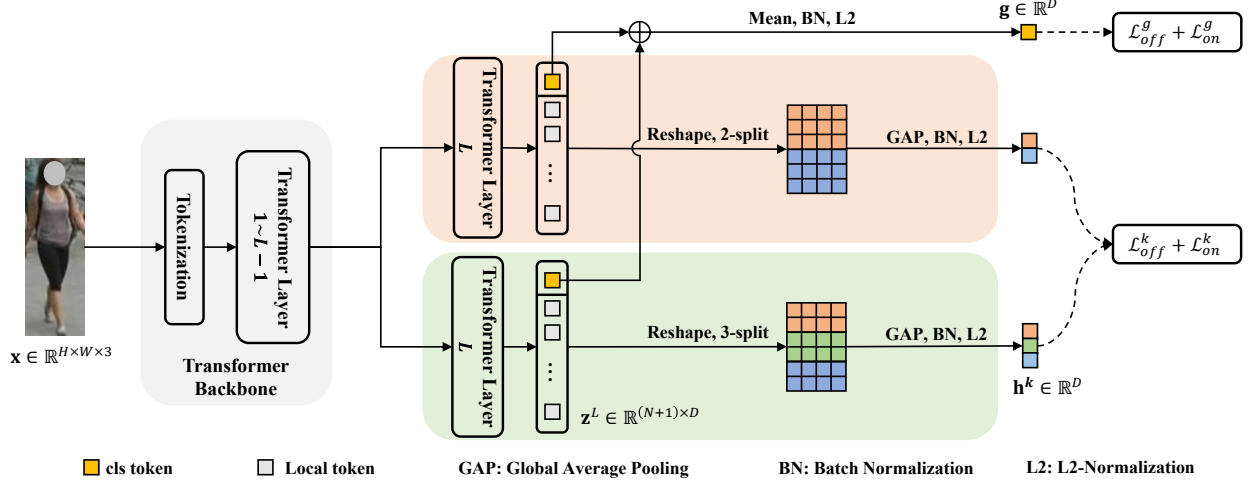
Figure 1. An overview of the proposed method. It constructs a dual-branch architecture appended to a pure transformer backbone to extract features at multiple granularities. Besides, offline and online contrastive learning losses defined with respect to both global and part-level features are employed for unsupervised learning.
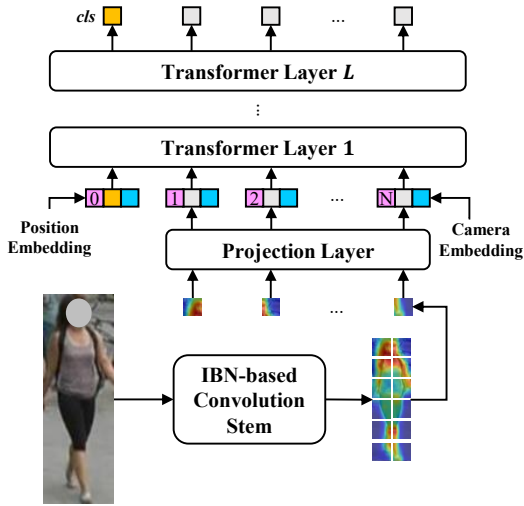

Figure 2. An illustration of the transformer backbone [28]. It builds upon ViT [9] but makes the following modifications: 1) replace the vanilla convolution stem [45] by the IBN-based convolution stem and 2) additionally include camera embeddings.

Besides, $\tilde{y}_i \in \{1, \cdots, N_p\}$ is a generated pseudo label, N_p is the number of proxies and N_I is the number of images. Then, a proxy-level memory bank $\mathcal{K} \in \mathbb{R}^{N_p \times D}$ is constructed, in which each entry stores the centroid of a proxy's global feature. During back-propagation, when image $\mathbf{x}_i$ is input, the entry corresponding to the pseudo class $\tilde{y}_i$ is updated via

$$\mathcal{K}[\tilde{y}_i] \leftarrow \mu \mathcal{K}[\tilde{y}_i] + (1 - \mu)\mathbf{g}_i, \qquad (6)$$

where $\mathcal{K}[\tilde{y}_i]$ denotes the $\tilde{y}_i$-th entry of the memory bank and

$\mu \in [0, 1]$ is an updating rate.

Assisted with the memory bank, two contrastive learning losses are designed based on offline and online associations respectively. For image $\mathbf{x}_i$, the offline association retrieves a positive proxy set $\mathcal{P}_1$ directly according to the offline clustering and splitting results while gets a negative set $\mathcal{Q}_1$ from the remaining hard negative proxies. $\mathcal{P}_1$ and $\mathcal{Q}_1$, respectively, store the indexes of associated positive and negative proxies. Then, the offline contrastive loss is defined as

$$\mathcal{L}_{off}^g = -\sum_{i=1}^{B} \left(\frac{1}{|\mathcal{P}_1|} \sum_{u \in \mathcal{P}_1} \log \frac{S(u, \mathbf{g}_i)}{\sum\limits_{p \in \mathcal{P}_1} S(p, \mathbf{g}_i) + \sum\limits_{q \in \mathcal{Q}_1} S(q, \mathbf{g}_i)} \right),$$

$$(7)$$

in which $S(u, \mathbf{g}_i) = \exp(\mathcal{K}[u]^T \mathbf{g}_i/\tau)$, τ is a temperature factor. Moreover, $|\cdot|$ denotes the cardinality of a set and B is the batch size.

However, offline association is noisy due to the imperfect clustering results. To remedy the noise, an online association strategy is proposed. It employs an instance-proxy balanced similarity and a camera-aware nearest neighbor scheme to select a positive proxy set $\mathcal{P}_2$ and a negative set $\mathcal{Q}_2$ for each anchor image $\mathbf{x}_i$ on the fly. The online contrastive loss is defined as

$$\mathcal{L}_{on}^g = -\sum_{i=1}^{B} \left(\frac{1}{|\mathcal{P}_2|} \sum_{u \in \mathcal{P}_2} \log \frac{S(u, \mathbf{g}_i)}{\sum\limits_{p \in \mathcal{P}_2} S(p, \mathbf{g}_i) + \sum\limits_{q \in \mathcal{Q}_2} S(q, \mathbf{g}_i)} \right).$$

$$(8)$$

The above-introduced losses are defined with respect to global features. In order to leverage part-level features, we additionally construct memory banks and define both types of losses for each part, while keeping the clustering step unchanged (i.e. only using the global features for clustering

as in O2CAP [39]). The entire loss for training is as follows:

$$\mathcal{L} = \mathcal{L}_{off}^g + \mathcal{L}_{on}^g + \lambda_p \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{L}_{off}^k + \mathcal{L}_{on}^k \right), \quad (9)$$

where λ_p is a weighting factor to balance global and part-based losses.

4. Experiments

4.1. Datasets and Evaluation Metrics

To validate the proposed method, we conduct a series of experiments on three person Re-ID datasets: Market1501 [54], DukeMTMC-reID [31], and MSMT17 [41]. The former two datasets are captured on university campus with outdoor scenarios only, while MSMT17 [41] contains both indoor and outdoor scenarios and therefore is more challenging. The number of cameras, together with the number of IDs and images contained in training, query, and gallery sets on three datasets are listed in Table 1. The training sets are used for unsupervised learning. During test time, each image in query is matched to similar images in gallery sets.

Dataset	Training Set			Query Set		Gallery Set	
	#Camera	#ID	#Image	#ID	#Image	#ID	#Image
Market1501	6	751	12,936	750	3,368	751	15,913
DukeMTMC-reID	8	702	16,522	702	2,228	1,110	17,661
MSMT17	15	1,041	32,621	3,060	11,659	3,060	82,161

Table 1. The number of cameras, IDs and images on three datasets.

As the common practice [3, 39], we employ the extensively used mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) for performance evaluation. The CMC metric is reported via Rank-1, Rank-5, and Rank-10. Besides, no post-processing techniques (e.g. Re-ranking [57]) are used.

4.2. Implementation Details

Our transformer backbone is built upon the ViT-Small/16 model [9, 28], which has $L = 12$ transformer layers and the number of heads in each MSA is 6, the feature dimension is $D = 384$. The backbone is pre-trained on a large-scale unlabeled dataset LUPerson [12]. Each image is resized to 384×128 and augmented with random horizontal flipping, cropping and erasing [56]. Patch size $P = 16$, that is, each feature map produced after the IBN-based convolution stem is split into patches in size of 8×8. The weight for camera embedding is $\lambda_c = 3$. The numbers of parts in two branches are $K_1 = 2$ and $K_2 = 3$ unless otherwise specified.

In unsupervised learning, all involved hyper-parameters are set the same as those in O2CAP [39]. That is, the updating rate $\mu = 0.2$, the temperature factor $\tau = 0.07$, and the batch size $B = 32$. Moreover, the weight of part-based losses is $\lambda_p = 0.1$. The model is trained by SGD optimizer

for 50 epochs, with a momentum of 0.9, a learning rate of 0.00035 and a weight decay of 0.0005. The learning rate is regulated by a warmup scheduler that multiplies 0.01 on the initial learning rate and linearly enlarges it to the standard value in previous 10 epochs. At epoch 20 and 40, the learning rate is divided by 10. The entire algorithm is implemented in PyTorch [30]. To accelerate training, we use half-precision floating point (FP16) computation in training while use full precision (FP32) for test. All experiments are run on a single NVIDIA GTX 1080 GPU.

4.3. Ablation Studies

We first conduct a series of experiments to validate the effectiveness of the proposed method, which is referred to as the Transformer-based Multi-Grained Feature (TMGF) method. Different variants of our model are investigated and all experiments are conducted on MSMT17 [41].

Effectiveness of Two Branches. In order to investigate the effectiveness of our dual-branch architecture, we compare the full model with three variants, including: 1) TMGF1: a baseline model that simply adopts the original transformer backbone and uses the global token as the image feature for learning; 2) TMGF2: a model with a single branch and the local tokens are partitioned into two parts, and meanwhile both global and part-level features are taken for learning; and 3) TMGF3: a model with a single branch similar to TMGF2 but with three partitions for local tokens. The performance of all model variants are reported in Table 2. From the results we observe that the leverage of local tokens can consistently improve the performance, and 2-partition outperforms 3-partition if only a single branch is constructed. When two branches are working together, the mAP performance is further boosted while Rank-1 is slightly dropped (comparing TMGF vs. TMGF2).

Models	Duplicate Layer L	Partition		MSMT17	
		2-split	3-split	mAP	Rank-1
TMGF1 (Baseline)				53.5	81.7
TMGF2 (Single B.)		✓		57.8	**83.6**
TMGF3 (Single B.)			✓	54.7	82.0
TMGF4 (w/o Dup.)		✓	✓	56.8	82.8
TMGF (Full)	✓	✓	✓	**58.2**	83.3

Table 2. Comparison of the proposed full model and its variants in different architectures. "Single B." denotes "single branch" and "w/o Dup." denotes "without duplication".

Effectiveness of Duplicating Layer L. In our dual-branch architecture, the last transformer layer (i.e. layer L) is duplicated so that each branch has one copy to produce independent outputs for different partition. An alternative design is to simply append two branches after the original transformer backbone without duplication, which performs two

different partitions on the same output local tokens. We refer to this design as the TMGF4 model, whose performance is also presented in Table 2. When different partitions are conducted on the same outputs, the network may mix up the learning of local cues, leading to a degenerated performance. Therefore, it is necessary to duplicate the last transformer layer for better learning.

Impact of Feature Granularity. The granularity of part-level features may influence the performance as well. Thus, we investigate three different partition ways, which split local tokens of two branches into 2 and 3 parts, or 2 and 4 parts, or 3 and 4 parts respectively. Table 3 lists the performance under different partitions. We observe that the best performance is achieved by the one with 2 and 3 parts in two branches. Splitting local tokens into more parts does not bring improvement. On the contrary, more fine-grained partition degenerates the performance.

Granularities	MSMT17	
	mAP	Rank-1
$[2, 3]$	**58.2**	**83.3**
$[2, 4]$	57.1	82.9
$[3, 4]$	56.6	82.7

Table 3. The performance under different partitions. $[K_1, K_2]$ denotes that the local tokens in the first and second branches are split into K_1 and K_2 parts respectively.

Impact of the Fusion of Global Tokens. Each branch of our dual-branch architecture yields a global token. It implies that there are different ways to generate the global feature for learning. For example, we can choose the global token produced in either branch as the global feature, or average both global tokens as introduced in Section 3.2. In this experiment we investigate the different ways and compare their performance in Table 4. The results show that the way of averaging both global tokens outperforms the other ways by a considerable margin, because it encourages a more balanced learning of two branches.

Models	Global Feature	MSMT17	
		mAP	Rank-1
TMGF5	$\mathbf{g} = cls_1^L$	55.9	82.5
TMGF6	$\mathbf{g} = cls_2^L$	56.8	82.8
TMGF(Full)	$\mathbf{g} = \frac{1}{2}\left(cls_1^L + cls_2^L\right)$	**58.2**	**83.3**

Table 4. The performance of the models using different global features. cls_1^L and cls_2^L denote the global tokens output in the first and second branch respectively.

4.4. Visualization of Attention Maps

To further illustrate the effectiveness of our dual-branch architecture, we visualize attention maps via the Attention Rollout scheme [1]. Figure 3 presents the attention maps obtained by the full model (TMGF), together with the maps obtained by the baseline model (TMGF1) for comparison. We notice that, with the exploitation of multi-grained features, our full model can attend to person regions more completely (e.g. examples in the first row) and focus more on discriminative regions (e.g. examples in the fourth row). Meanwhile, the full model is less distracted by background (e.g. examples in the second row) or occluded regions (e.g. examples in the third row). These properties enable our model to achieve a promising Re-ID performance even if the dataset is very challenging.

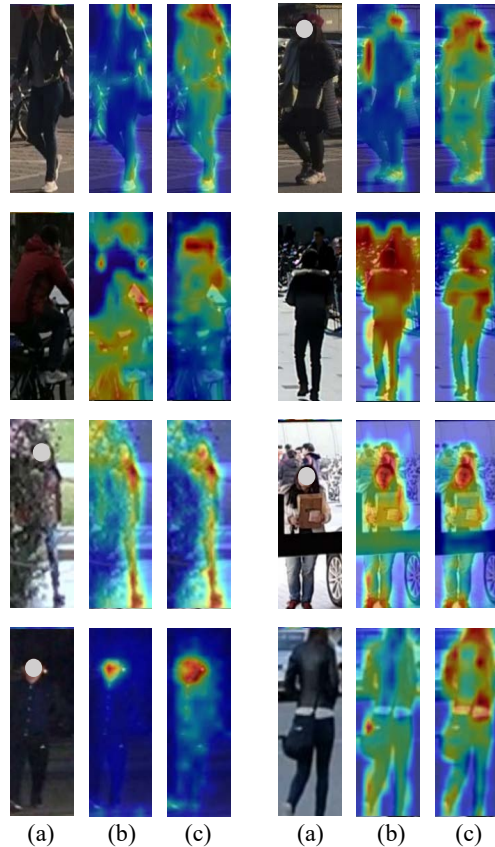|					
(a)	(b)	(c)	(a)	(b)	(c)

Figure 3. Visualization of attention maps. (a) color images, (b) attention maps obtained by the baseline model (TMGF1), and (c) attention maps obtained by our full model (TMGF).

4.5. Comparison to State-of-the-arts

Finally, we compare our approach with state-of-the-art methods on three person Re-ID datasets. Table 5 presents the comparison results.

Methods	Reference	Market1501				DukeMTMC-reID				MSMT17			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Purely unsupervised methods													
BUC [24]	AAAI19	38.3	66.2	79.6	84.5	27.5	47.4	62.6	68.4	-	-	-	-
HCT [49]	CVPR20	56.4	80.0	91.6	95.2	50.7	69.6	83.4	87.4	-	-	-	-
ClusterContrast [8]	arxiv21	83.0	92.9	97.2	98.0	73.6	85.5	92.2	94.3	31.2	61.5	71.8	76.7
CAP [38]	AAAI21	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8	36.9	67.4	78.0	81.4
IICS [46]	CVPR21	72.9	89.5	95.2	97.0	64.4	80.0	89.0	91.6	26.9	56.4	68.8	73.4
RLCC [51]	CVPR21	77.7	90.8	96.3	97.5	69.2	83.2	91.6	93.8	27.9	56.5	68.4	73.1
ICE [3]	ICCV21	82.3	93.8	97.6	98.4	69.9	83.3	91.5	94.1	38.9	70.2	80.5	84.4
MGH [43]	MM21	81.7	93.2	96.8	98.1	70.2	83.7	92.1	93.7	40.6	70.2	81.2	84.5
MCRN [42]	AAAI22	80.8	92.5	-	-	69.9	83.5	-	-	31.2	63.6	-	-
MGCE-HCL [34]	ACPR22	79.6	92.1	-	-	67.5	82.5	-	-	-	-	-	-
O2CAP [39]	arxiv22	82.7	92.5	96.9	98.0	71.2	83.9	91.3	93.4	42.4	72.0	81.9	85.4
PPLR [7]	CVPR22	84.4	94.3	97.8	98.6	-	-	-	-	42.2	73.3	83.5	86.5
TransReID-SSL† [28]	arxiv21	**89.6**	95.3	-	-	-	-	-	-	50.6	75.0	-	-
TMGF†	This work	89.5	**95.5**	**98.0**	**98.7**	**76.8**	**86.7**	**92.9**	**94.1**	**58.2**	**83.3**	**90.2**	**92.1**
UDA-based methods													
SSG [13]	ICCV19	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2	13.3	32.2	-	51.2
SpCL [14]	NIPS20	76.7	90.3	96.2	97.7	68.8	82.9	90.1	92.5	26.5	53.1	65.8	70.5
Isobe *et al.* [19]	ICCV21	83.4	94.2	-	-	70.8	83.5	-	-	36.3	66.6	-	-
DARC [18]	AAAI22	85.1	94.1	97.6	98.7	-	-	-	-	35.2	64.5	76.2	80.4
Fully supervised methods													
ABD-Net [4]	ICCV19	88.3	95.6	-	-	78.6	89.0	-	-	60.8	82.3	90.6	-
st-ReID [36]	AAAI19	86.7	97.2	99.3	99.5	82.8	94.0	97.0	97.8	-	-	-	-
PAT† [23]	CVPR21	88.0	95.4	-	-	78.2	88.8	-	-	-	-	-	-
TransReID† [17]	ICCV21	89.5	95.2	-	-	82.6	90.7	-	-	69.4	86.2	-	-
LA-Transformer† [32]	arxiv21	94.5	98.3	-	-	-	-	-	-	-	-	-	-
PFD† [40]	AAAI22	89.7	95.5	-	-	83.2	91.2	-	-	-	-	-	-
TransReID-SSL(w/ GT)† [28]	arxiv21	91.3	96.2	-	-	-	-	-	-	68.1	86.1	-	-
TMGF(w/ GT)†	This work	91.9	96.3	98.9	99.3	83.1	92.3	96.4	97.4	70.3	88.2	94.1	95.4

Table 5. Comparison with state-of-the-art methods. † indicates that the method is using a transformer backbone.

Comparison with Unsupervised Methods. We include 17 representative or recent unsupervised person Re-ID methods for comparison, among which 13 methods are purely unsupervised and 4 methods are UDA-based. Except TransReID-SSL [28], all previous methods are based on CNN backbones pre-trained on ImageNet [20]. It can be seen that the transformer feature extraction backbone [28] greatly boosts the performance because of its network architecture and the pre-train on LUPerson [12]. Our method is built upon TransReID-SSL [28]. In contrast to TransReID-SSL that uses contrastive learning losses defined in ClusterContrast [8], we design a dual-branch architecture to extract multi-grained features and define both global and part-based contrastive learning losses in the form of O2CAP [39]. On the most challenging dataset MSMT17, our method surpasses TransReID-SSL by 7.6% mAP and 8.3% Rank-1. When compared to the CNN-based O2CAP method, we improve the performance by a significant margin. Especially on MSMT17, 15.8% mAP and 11.3% Rank-1 improvements are achieved.

Comparison with Fully Supervised Methods. We include seven recent fully supervised Re-ID methods for reference, in which ABD-Net [4] and st-ReID [36] are CNN-based and all others [17, 23, 28, 32, 40] are transformer-based. Meanwhile, the performance of our model trained with ground-

truth labels is also provided, which indicates the upper bound performance our method can achieve. From the results we see that our unsupervised model has already performed better than several supervised methods on Market-1501. The performance gap between our unsupervised model and the supervised ABD-Net [4] is very small on DukeMTMC-reID and MSMT17. In addition, our model trained with ground-truth performs better than the supervised TransReID-SSL [36], validating the effectiveness of our multi-grained feature extraction.

5. Conclusion

In this work, we have presented an approach to extract multi-grained features from a pure transformer network and leverage the multi-grained features for unsupervised person Re-ID. The designed dual-branch architecture is simple but effective for feature extraction. Benefited from the multi-grained features and part-feature based learning losses, our method outperforms existing unsupervised Re-ID methods by a considerable margin, greatly mitigating the performance gap to supervised counterparts.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, pages 14960–14969, 2021.

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[7] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7308–7318, 2022.

[8] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Xiaoli Liu, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. *arXiv preprint arXiv:2103.11568*, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[11] Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, 30:2898–2907, 2021.

[12] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021.

[13] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, pages 6112–6121, 2019.

[14] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, pages 11309–11321, 2020.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

[18] Zhengdong Hu, Yifan Sun, Yi Yang, and Jianguang Zhou. Divide-and-regroup clustering for domain adaptive person re-identification. In *AAAI*, 2022.

[19] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *ICCV*, pages 8526–8536, 2021.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[21] Shenqi Lai, Zhenhua Chai, and Xiaolin Wei. Transformer meets part model: Adaptive part division for person re-identification. In *ICCVW*, pages 4150–4157, 2021.

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018.

[23] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021.

[24] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, pages 8738–8745, 2019.

[25] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In *CVPR*, pages 3390–3399, 2020.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[27] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019.

[28] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

[29] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[32] Charu Sharma, Siddhant R Kapil, and David Chapman. Person re-identification with a locally aware transformer. *arXiv preprint arXiv:2106.03720*, 2021.

[33] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, pages 5363–5372, 2018.

[34] He Sun, Mingkun Li, and Chun-Guang Li. Hybrid contrastive learning with cluster ensemble for unsupervised person re-identification. In *ACPR*, pages 532–546. Springer, 2022.

[35] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[36] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *AAAI*, volume 33, pages 8933–8940, 2019.

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018.

[38] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, pages 2764–2772, 2021.

[39] Menglin Wang, Jiachen Li, Baisheng Lai, Xiaojin Gong, and Xian-Sheng Hua. Offline-online associated camera-aware proxies for unsupervised person re-identification. *arXiv preprint arXiv:2201.05820*, 2022.

[40] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *AAAI*, pages 2540–2549, 2022.

[41] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[42] Yuhang Wu, Tengteng Huang, Haotian Yao, Chi Zhang, Yuanjie Shao, Chuchu Han, Changxin Gao, and Nong Sang. Multi-centroid representation network for domain adaptive person re-id. In *AAAI*, pages 2750–2758, 2022.

[43] Yiming Wu, Xintian Wu, Xi Li, and Jian Tian. Mgh: Metadata guided hypergraph modeling for unsupervised person re-identification. In *ACM MM*, pages 1571–1580, 2021.

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[45] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *NeurIPS*, 34:30392–30400, 2021.

[46] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021.

[47] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, pages 3633–3642, 2019.

[48] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *CVPR*, pages 14298–14307, 2022.

[49] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, pages 13657–13665, 2020.

[50] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, pages 516–525, 2021.

[51] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *CVPR*, pages 3436–3445, 2021.

[52] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019.

[53] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, pages 5310–5319, 2021.

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[55] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *ICCV*, pages 8371–8381, 2021.

[56] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.

[57] Zhun Zhong, Liang Zheng, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[58] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021.