TOP-ReID: Multi-Spectral Object Re-identification with Token Permutation

Yuhao Wang[1], Xuehu Liu[2], Pingping Zhang[1]*, Hu Lu[3], Zhengzheng Tu[4], Huchuan Lu[1]

[1]School of Future Technology, School of Artificial Intelligence, Dalian University of Technology
[2]School of Computer Science and Artificial Intelligence, Wuhan University of Technology
[3]School of Computer Science and Communication Engineering, Jiangsu University
[4]School of Computer Science and Technology, Anhui University
924973292@mail.dlut.edu.cn, liuxuehu@whut.edu.cn, {zhpp, lhchuan}@dlut.edu.cn, luhu@ujs.edu.cn
zhengzhengahu@163.com

Abstract

Multi-spectral object Re-identification (ReID) aims to retrieve specific objects by leveraging complementary information from different image spectra. It delivers great advantages over traditional single-spectral ReID in complex visual environment. However, the significant distribution gap among different image spectra poses great challenges for effective multi-spectral feature representations. In addition, most of current Transformer-based ReID methods only utilize the global feature of class tokens to achieve the holistic retrieval, ignoring the local discriminative ones. To address the above issues, we step further to utilize all the tokens of Transformers and propose a cyclic token permutation framework for multi-spectral object ReID, dubbled TOP-ReID. More specifically, we first deploy a multi-stream deep network based on vision Transformers to preserve distinct information from different image spectra. Then, we propose a Token Permutation Module (TPM) for cyclic multi-spectral feature aggregation. It not only facilitates the spatial feature alignment across different image spectra, but also allows the class token of each spectrum to perceive the local details of other spectra. Meanwhile, we propose a Complementary Reconstruction Module (CRM), which introduces dense token-level reconstruction constraints to reduce the distribution gap across different image spectra. With the above modules, our proposed framework can generate more discriminative multi-spectral features for robust object ReID. Extensive experiments on three ReID benchmarks (i.e., RGBNT201, RGBNT100 and MSVR310) verify the effectiveness of our methods. The code is available at https://github.com/924973292/TOP-ReID.

Introduction

Object Re-identification (ReID) aims to retrieve specific objects from images or videos across non-overlapping cameras, which has advanced significantly over the past decades. In the traditional object ReID, researchers primarily utilize single-spectral images (such as RGB, depth) to extract visual information of the targets. However, single-spectral images provide very limited representation abilities in scenarios characterized by low resolution, darkness, glare, etc. As illustrated in the top row of Fig. 1, the outlines of persons are notably blurred, leading to an evident confusion between persons and the background in the RGB image spectrum. Hence, relying only on RGB images poses great challenges for robust object ReID. Fortunately, other image spectra are very useful to address above problems. In fact, Near Infrared (NIR) imaging is unaffected by darkness and adverse weather conditions (Li et al. 2020b). Thus, there have been some efforts (Li et al. 2020a; Liu et al. 2021a; Zhang and Wang 2023) to incorporate NIR images to enhance the performance of object ReID. Nonetheless, NIR images retain some limitations (Zheng et al. 2021), as depicted in Fig. 1. For example, the details of persons in NIR images tend to be substantially obscured in the presence of glare. Meanwhile, Thermal Infrared (TIR) imaging is more robust to these scenarios (Zheng et al. 2021). As illustrated in Fig. 1, TIR images can highlight persons from the background and preserve crucial details, such as glasses and backpacks. These facts clearly show the information complementarity of different image spectra for object ReID. Based on the above facts, multi-spectral object ReID aims to retrieve specific ob-

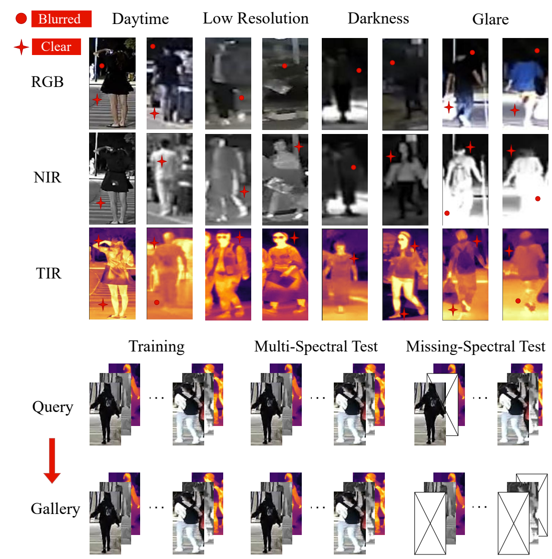


Figure 1: The top displays instances from RGBNT201 in various challenges, while the bottom presents the object ReID settings for multi-spectral and missing-spectral test.

*Corresponding author

jects by leveraging complementary information from different image spectra, e.g., RGB, NIR, TIR etc. It delivers great advantages over single-spectral ReID in complex visual environment. In fact, some methods (Zheng et al. 2021; Wang et al. 2022b) have already tried to integrate multi-spectral features with simple fusion methods. However, there are significant distribution gaps among different image spectra. Simple fusions can not well address the heterogenous challenges for effective feature representations. In addition, it often involves the absence of image spectra in real world, as shown in the "Missing-spectral Test" of Fig. 1. Thus, there are much room for improving multi-spectral feature fusion.

Meanwhile, with the great advance of vision Transformers (Dosovitskiy et al. 2020), some works (He et al. 2021; Pan et al. 2022) have introduced Transformers for object ReID. However, most of current Transformer-based ReID methods only utilize the global feature of class tokens to achieve the holistic retrieval, ignoring the local discriminative ones. To address the above issues, we step further to utilize all the tokens of Transformers and propose a cyclic token permutation framework for multi-spectral object ReID, dubbled TOP-ReID. Specifically, it consists of two key modules: Token Permutation Module (TPM) and Complementary Reconstruction Module (CRM). Technically, we first deploy a multi-stream deep network based on vision Transformers to preserve distinct information from different image spectra. Then, TPM takes all the tokens from the multi-stream deep network as inputs, and cyclically permutes the specific class tokens and the corresponding patch tokens from other spectra. In this way, it not only facilitates the spatial feature alignment across different image spectra, but also allows the class token of each spectrum to perceive the local details of other spectra. Meanwhile, CRM is proposed to facilitate local information interaction and reconstruction across different image spectra. Through introducing token-level reconstruction constraints, it can reduce the distribution gap across different image spectra. As a result, the CRM can further handle the missing-spectral problem. With the proposed modules, our framework can extract more discriminative features from multi-spectral images for robust object ReID. Comprehensive experiments are conducted on three multi-spectral object ReID benchmarks, i.e., RGBNT201, RGBNT100 and MSVR310. Experimental results clearly show the effectiveness of our proposed methods.

In summary, our contributions can be stated as follows:

- We propose a novel feature learning framework named TOP-ReID for multi-spectral object ReID. To our best knowledge, our proposed TOP-ReID is the first work to utilize all the tokens of vision Transformers to improve the multi-spectral object ReID.
- We propose a Token Permutation Module (TPM) and a Complementary Reconstruction Module (CRM) to facilitate multi-spectral feature alignment and handle spectral-missing problems effectively.
- We perform comprehensive experiments on three multi-spectral object ReID benchmarks, i.e., RGBNT201, RGBNT100 and MSVR310. The results fully verify the effectiveness of our proposed methods.

Related Work

Single-spectral Object ReID

Single-spectral object ReID focuses on extracting discriminative features from single-spectral images. Typical single-spectral forms include RGB, NIR, TIR and depth. Due to the easy requirement, RGB images play a fundamental role in the single-spectral object ReID. As for the techniques, most of existing object ReID methods are based on Convolutional Neural Networks (CNNs). For example, Luo *et al.* (Luo et al. 2019) utilize a deep residual network and introduce the BNNeck technique for object ReID. Furthermore, PCB (Sun et al. 2018) and MGN (Wang et al. 2018) adapt a stripe-based image division strategy to obtain multi-grained representations. OSNet (Zhou et al. 2019) employs a unified aggregation gate for fusing omni-scale features. AGW (Ye et al. 2021) incorporates non-local attention mechanisms for fine-grained feature extraction. Nevertheless, due to the limited receptive field, CNN-based methods(Qian et al. 2017; Li, Zhu, and Gong 2018; Chang, Hospedales, and Xiang 2018; Chen et al. 2019; Sun et al. 2020; Rao et al. 2021; Zhao et al. 2021; Liu et al. 2021b) are not robust to complex scenarios. Inspired by the success of vision Transformers (ViT) (Dosovitskiy et al. 2020), He *et al.* (He et al. 2021) propose the first pure Transformer-based method named TransReID for object ReID, yielding competitive results through the adaptive modeling of image patches. Afterwards, numerous Transformer-based methods (Zhu et al. 2021; Zhang et al. 2021; Chen et al. 2022; Wang et al. 2022a; Liu et al. 2023) demonstrate their advantages in object ReID. However, all these methods take single-spectral images as inputs, providing limited representation abilities. Thus, they can not handle the all-day object ReID problem.

Multi-spectral Object ReID

The robustness of multi-spectral data draws the attention of numerous researchers. For multi-spectral person ReID, Zheng *et al.* (Zheng et al. 2021) advance the field and design a PFNet to learn robust RGB-NIR-TIR features. Then, Wang *et al.* (Wang et al. 2022b) boost modality-specific representations with three learning strategies, named IEEE. Furthermore, Zheng *et al.* (Zheng et al. 2023) design a DENet to address the spectral-missing problem. For multi-spectral vehicle ReID, Li *et al.* (Li et al. 2020b) propose a HAMNet to fuse different spectral features. Considering the relationship between different image spectra, Guo *et al.* (Guo et al. 2022) propose a GAFNet to fuse the multiple data sources. He *et al.* (He et al. 2023) propose a GPFNet to adaptively fuse multi-spectral features. Zheng *et al.* (Zheng et al. 2022) propose a CCNet to simultaneously overcome the discrepancies from both modality and sample aspects. Pan *et al.* (Pan et al. 2022) propose a HViT to balance modal-specific and modal-shared information. Furthermore, they employ a random hybrid augmentation and a feature hybrid mechanism to improve the performance (Pan et al. 2023). Although effective, previous methods mainly treat the NIR and TIR as an assistant to RGB, rather than adaptively fuse them with multi-level spatial correspondences. In contrast, we facilitate the spatial feature alignment across different image spectra.

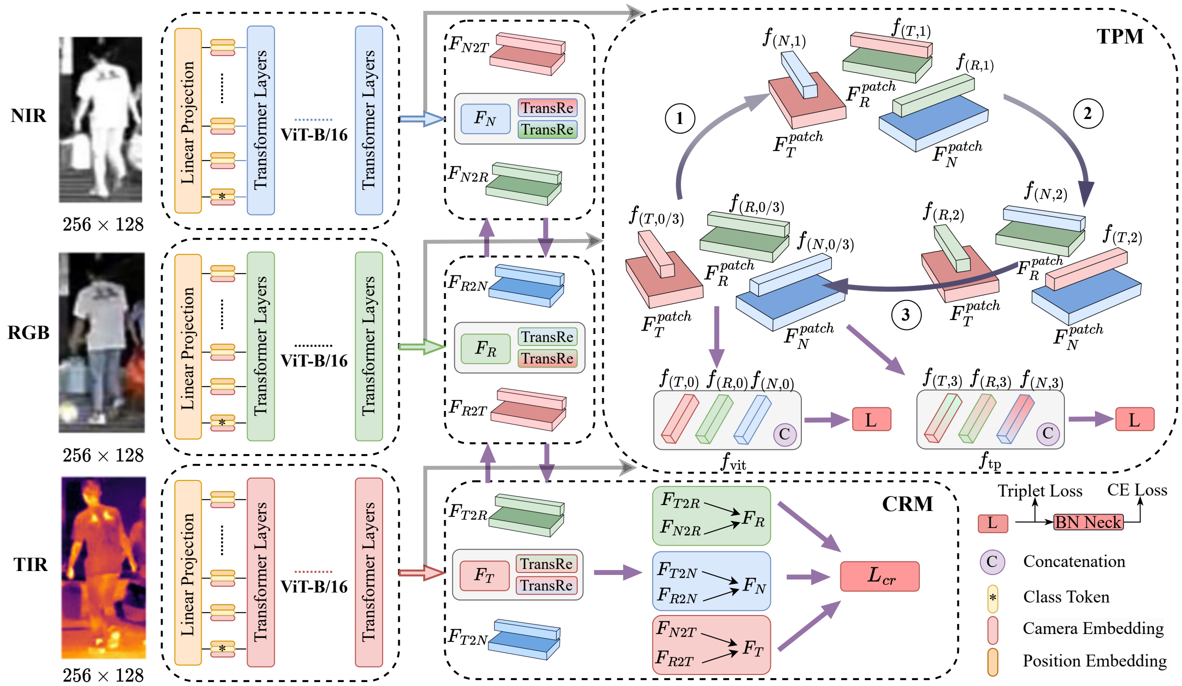


Figure 2: An illustration of the proposed TOP-ReID. First, deep features from RGB, NIR and TIR images are extracted by using three independent ViT-B/16. Then, a Token Permutation Module (TPM) is proposed for cyclic multi-spectral feature aggregation through three consecutive token permutations. Meanwhile, a Complementary Reconstruction Module (CRM) is used to achieve token-level reconstruction constraints. When inference, we utilize the permutated features for ranking the person candidates.

Proposed Method

As illustrated in Fig. 2, our proposed TOP-ReID consists of three main components: Multi-stream Feature Extraction, Token Permutation Module (TPM) and Complementary Reconstruction Module (CRM).

Multi-stream Feature Extraction

In this work, we take images of three spectra for object ReID, i.e., RGB, NIR and TIR. To capture the distinctive characteristics of each spectrum, we follow previous works (Li et al. 2020b; Zheng et al. 2021) and adopt three independent backbones. More specifically, vision Transformers (ViT) can be deployed as the backbone in each stream. Formally, the multi-stream features can be represented as

$$F_{\mathrm{R}} = \mathrm{ViT}_{\mathrm{R}}\left(I_{\mathrm{R}}\right), \tag{1}$$

$$F_{\mathrm{N}} = \mathrm{ViT}_{\mathrm{N}}\left(I_{\mathrm{N}}\right), \tag{2}$$

$$F_{\mathrm{T}} = \mathrm{ViT}_{\mathrm{T}}\left(I_{\mathrm{T}}\right), \tag{3}$$

where $I_{\mathrm{R}} \in R^{H \times W \times 3}$, $I_{\mathrm{N}} \in R^{H \times W \times 3}$ and $I_{\mathrm{T}} \in R^{H \times W \times 3}$ denote the input RGB, NIR and TIR images, respectively. Here, ViT can be any vision Transformers (e.g., ViT-B/16 (Dosovitskiy et al. 2020), DeiT-S/16 (Touvron et al. 2021), T2T-ViT-24 (Yuan et al. 2021)). The token features F_{R}, F_{N}, $F_{\mathrm{T}} \in R^{D \times (M+1)}$ are extracted from the final layer of ViT, respectively. Additional learnable class token is included. D denotes the embedding dimension while M means the number of patch tokens. These independent streams enable the extraction of spectral-specific features, capturing rich information from different image spectra.

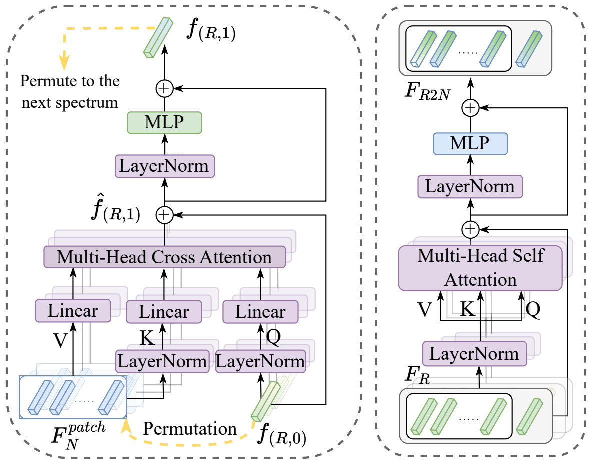
Token Permutation Module

To achieve the spatial feature alignment among different image spectra and the effective aggregation of heterogeneous features, we introduce the Token Permutation Module (TPM) with a cyclic token permutation mechanism, as illustrated at the top right corner of Fig. 2.

Technically, TPM takes the token features F_{R}, F_{N} and F_{T} as inputs, and generates the fused feature f_{tp} with three consecutive token permutations. Without loss of generality, we take the RGB stream as a starting example. As shown in Fig. 3 (a), we utilize a Multi-Head Cross-Attention (MHCA) (Dosovitskiy et al. 2020) with N_h heads to achieve the token permutation. More specifically, the class token $f_{(\mathrm{R},0)} \in R^{D}$ from F_{R} is passed into a linear transformation to generate a query matrix $Q \in R^{D}$. The patch tokens $F_{\mathrm{N}}^{patch} \in R^{D \times M}$ from F_{N} are passed into two linear transformations to generate a key matrix K and a value matrix V, respectively. Thus, the interaction of F_{R} and F_{N} in the h-th head is represented as

$$\hat{f}^{h}_{(\mathrm{R},1)} = \sigma\left(\frac{Q^h {K^h}^{\top}}{\sqrt{d}}\right) V^h, \tag{4}$$

where σ is the softmax function and $(\cdot)^{\top}$ means the matrix transposition. Here, $Q^h \in R^{d}$, K^h, $V^h \in R^{d \times M}$, $d = \frac{D}{N_h}$.



(a) Token Permutation (b) TransRe Block

Figure 3: Our token permutation and TransRe blocks with the RGB stream. Other streams share a similar structure.

The outputs of N_h heads $(\hat{f}^1_{(\mathrm{R},1)}, \cdots, \hat{f}^h_{(\mathrm{R},1)}, \cdots, \hat{f}^{N_h}_{(\mathrm{R},1)})$ are concatenated to be $\hat{f}_{(\mathrm{R},1)} \in R^D$. Then, $\hat{f}_{(\mathrm{R},1)}$ is passed through a Feed-Forward Network (FFN) to generate a new class token $f_{(\mathrm{R},1)}$,

$$f_{(\mathrm{R},1)} = \mathrm{FFN}(\hat{f}_{(\mathrm{R},1)}) + \hat{f}_{(\mathrm{R},1)}. \tag{5}$$

It serves as the initial spatial alignment of the RGB and NIR image features. Similar operations can be performed for other spectra,

$$f_{(\mathrm{R},1)} = \mathrm{FFN}(\mathrm{MHCA}\left(\mathrm{LN}(f_{(\mathrm{R},0)}), \mathrm{LN}(F_{\mathrm{N}}^{patch})\right)), \tag{6}$$

$$f_{(\mathrm{N},1)} = \mathrm{FFN}(\mathrm{MHCA}\left(\mathrm{LN}(f_{(\mathrm{N},0)}), \mathrm{LN}(F_{\mathrm{T}}^{patch})\right)), \tag{7}$$

$$f_{(\mathrm{T},1)} = \mathrm{FFN}(\mathrm{MHCA}\left(\mathrm{LN}(f_{(\mathrm{T},0)}), \mathrm{LN}(F_{\mathrm{R}}^{patch})\right)). \tag{8}$$

As shown in Fig. 3 (a) and above equations, we additionally introduce the LayerNorm (LN) (Ba, Kiros, and Hinton 2016) to Q and K to ensure the numerical stability. Thus, the first token permutation can be totally formulated as

$$f_{(\mathrm{R},1)}, f_{(\mathrm{N},1)}, f_{(\mathrm{T},1)} = \mathrm{TPM}^1(F_{\mathrm{R}}, F_{\mathrm{N}}, F_{\mathrm{T}}), \tag{9}$$

From the above equations, it can be observed that the token permutation enables the global class token from each spectrum to interact with the local patch tokens of the next spectrum, achieving the initial feature fusion and alignment.

Furthermore, the permutated class tokens $f_{(\mathrm{R},1)}$, $f_{(\mathrm{N},1)}$, and $f_{(\mathrm{T},1)}$ are paired with their initial patch tokens to form $F_{\mathrm{R}\to\mathrm{N}}$, $F_{\mathrm{N}\to\mathrm{T}}$, and $F_{\mathrm{T}\to\mathrm{R}}$, respectively. The class tokens keep shifting to the next spectrum,

$$f_{(\mathrm{R},2)}, f_{(\mathrm{N},2)}, f_{(\mathrm{T},2)} = \mathrm{TPM}^2(F_{\mathrm{R}\to\mathrm{N}}, F_{\mathrm{N}\to\mathrm{T}}, F_{\mathrm{T}\to\mathrm{R}}). \tag{10}$$

At this stage, each spectrum has already incorporated detail information from other spectra. Similar to the previous step, the permutated class tokens $f_{(\mathrm{R},2)}$, $f_{(\mathrm{N},2)}$, and $f_{(\mathrm{T},2)}$ are paired with permutated patch tokens to form $F_{\mathrm{RN}\to\mathrm{T}}$, $F_{\mathrm{NT}\to\mathrm{R}}$, and $F_{\mathrm{TR}\to\mathrm{N}}$, respectively. Finally, the token permutation process ends with each class token interacting with its own patch tokens,

$$f_{(\mathrm{R},3)}, f_{(\mathrm{N},3)}, f_{(\mathrm{T},3)} = \mathrm{TPM}^3(F_{\mathrm{RN}\to\mathrm{T}}, F_{\mathrm{NT}\to\mathrm{R}}, F_{\mathrm{TR}\to\mathrm{N}}). \tag{11}$$

Through the above token permutation, the information from all other spectra is conveyed to the patch tokens through the class token, enabling robust feature alignment. Finally, we concatenate the permutated class tokens to obtain the permutated representation $f_{\mathrm{tp}} \in R^{3D}$,

$$f_{\mathrm{tp}} = \mathrm{Concat}\left(f_{(\mathrm{R},3)}, f_{(\mathrm{N},3)}, f_{(\mathrm{T},3)}\right). \tag{12}$$

This cyclic token permutation enhances the spatial fusion and implicit alignment of deep features across spectra, improving the ability of inter-spectral dependencies.

Complementary Reconstruction Module

There are significant distribution gaps among different image spectra. In addition, it often involves the absence of certain image spectra in real world. Inspired by the image generation (Zhu et al. 2017), we propose a Complementary Reconstruction Module (CRM) to reduce the distribution gap across different image spectra. The key is to incorporate dense token-level reconstruction constraints.

Without loss of generality, we take the RGB stream as an example and consider the NIR and TIR spectra missing. To reconstruct the missing tokens, we pass F_R through a Transformer-based Reconstruction (TransRe) block (See Fig. 3 (b)) and generate the corresponding tokens by

$$F_{\mathrm{R2N}} = \mathrm{TransRe}(F_{\mathrm{R}}), \tag{13}$$

$$F_{\mathrm{R2T}} = \mathrm{TransRe}(F_{\mathrm{R}}), \tag{14}$$

where $F_{\mathrm{R2N}}, F_{\mathrm{R2T}} \in R^{D\times(M+1)}$ are the reconstructed tokens. The reconstructed tokens F_{R2N} and F_{R2T} are constrained by the real token features F_{N} and F_{T} using the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\mathrm{R2N}} = \frac{1}{M+1}\sum_{i=1}^{M+1} ||F_{\mathrm{R2N}} - F_{\mathrm{N}}||_2^2, \tag{15}$$

$$\mathcal{L}_{\mathrm{R2T}} = \frac{1}{M+1}\sum_{i=1}^{M+1} ||F_{\mathrm{R2T}} - F_{\mathrm{T}}||_2^2, \tag{16}$$

$$\mathcal{L}_{\mathrm{R}} = \mathcal{L}_{\mathrm{R2N}} + \mathcal{L}_{\mathrm{R2T}}. \tag{17}$$

Through the above token-level reconstruction constraints, the distribution gap between RGB and other spectra is reduced. To improve the reconstruction ability, we introduce similar constraints to all the image spectra and achieve a multi-spectral complementary reconstruction. The complementary reconstruction loss $\mathcal{L}_{cr}$ can be expressed as the sum of the individual losses for each spectrum:

$$\mathcal{L}_{cr} = \mathcal{L}_{\mathrm{R}} + \mathcal{L}_{\mathrm{N}} + \mathcal{L}_{\mathrm{T}}. \tag{18}$$

By introducing token-level constraints, our CRM effectively reduces the distribution gap among different image spectra. Moreover, it can generate corresponding tokens of missing spectra, ensuring a unified learning framework even in scenarios where one or more spectra are absent.

Dynamic Cooperation Between CRM and TPM

In this work, we further introduce the dynamic cooperation between the CRM and TPM to handle the absence of any image spectrum. For example, when the RGB image spectrum is missing, the token features F_{N} and F_{T} will activate their reconstruction blocks to generate the corresponding RGB token features F_{N2R} and F_{T2R}, respectively. Then, the reconstructed RGB token features can be represented as

$$\bar{F}_{\mathrm{R}} = \frac{(F_{\mathrm{N2R}} + F_{\mathrm{T2R}})}{2}. \tag{19}$$

Then, $\bar{F}_{\mathrm{R}}$, F_{N} and F_{T} can be fed into the TPM to perform the token permutation as normal. Hence, our CRM can dynamically cooperate with TPM, ensuring that the missing spectrum can still participate in the permutation process.

Objective Function

As illustrated in Fig. 2, our objective function comprises three components: loss for the ViT backbone, loss for the token permutation and loss for the CRM. As for the ViT backbone and the token permutation, they are both supervised by the label smoothing cross-entropy loss (Szegedy et al. 2016) and triplet loss (Hermans, Beyer, and Leibe 2017). Finally, the total loss in our framework can be defined by

$$\mathcal{L}_{total} = \mathcal{L}_{tri}^{ViT} + \mathcal{L}_{ce}^{ViT} + \mathcal{L}_{tri}^{TP} + \mathcal{L}_{ce}^{TP} + \mathcal{L}_{cr}. \tag{20}$$

Experiments

Dataset and Evaluation Protocols

To evaluate the performance, we adopt three multi-spectral object ReID datasets. RGBNT201 (Zheng et al. 2021) is the first multi-spectral person ReID dataset with RGB, NIR and TIR spectra. RGBNT100 (Li et al. 2020b) is a large-scale multi-spectral vehicle ReID dataset. MSVR310 (Zheng et al. 2022) is a small-scale multi-spectral vehicle ReID dataset with more complex scenarios. Following previous works, we adopt the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K ($K = 1, 5, 10$) as our evaluation metrics.

Implementation Details

Our model is implemented with the PyTorch toolbox. We conduct experiments with one NVIDIA A800 GPU. We use pre-trained Transformers on the ImageNet classification dataset (Deng et al. 2009) as our backbones. All images are resized to 256×128×3 pixels. When training, random horizontal flipping, cropping and erasing (Zhong et al. 2020) are used as data augmentation. We set the mini-batch size to 128. Each mini-batch consists of 8 randomly selected object identities, and 16 images are sampled for each identity. We use the Stochastic Gradient Descent (SGD) optimizer with a momentum coefficient of 0.9 and a weight decay of 0.0001. Furthermore, the learning rate is initialized as 0.009. The warmup strategy and cosine decay are used during training.

Methods		RGBNT201			
		mAP	R-1	R-5	R-10
Single	HACNN	21.3	19.0	34.1	42.8
	MUDeep	23.8	19.7	33.1	44.3
	OSNet	25.4	22.3	35.1	44.7
	MLFN	26.1	24.2	35.9	44.1
	CAL	27.6	24.3	36.5	45.7
	PCB	32.8	28.1	37.4	46.9
Multi	HAMNet	27.7	26.3	41.5	51.7
	PFNet	38.5	38.9	52.0	58.4
	DENet	42.4	42.2	55.3	<u>64.5</u>
	IEEE	<u>47.5</u>	<u>44.4</u>	<u>57.1</u>	63.6
	TOP-ReID*	**72.3**	**76.6**	**84.7**	**89.4**

Table 1: Performance comparison on RGBNT201. The best and second results are in bold and underlined, respectively. * signifies Transformer-based approaches, while others are CNN-based ones.

Methods		RGBNT100		MSVR310	
		mAP	R-1	mAP	R-1
Single	DMML	58.5	82.0	19.1	31.1
	Circle Loss	59.4	81.7	22.7	34.2
	PCB	57.2	83.5	23.2	42.9
	MGN	58.1	83.1	26.2	44.3
	BoT	78.0	95.1	23.5	38.4
	HRCN	67.1	91.8	23.4	44.2
	OSNet	75.0	95.6	28.7	44.8
	AGW	73.1	92.7	28.9	<u>46.9</u>
	TransReID*	75.6	92.9	18.4	29.6
Multi	GAFNet	74.4	93.4	-	-
	GPFNet	75.0	94.5	-	-
	PHT*	<u>79.9</u>	92.7	-	-
	PFNet	68.1	94.1	23.5	37.4
	HAMNet	74.5	93.3	27.1	42.3
	CCNet	77.2	<u>96.3</u>	**36.4**	**55.2**
	TOP-ReID*	**81.2**	**96.4**	<u>35.9</u>	44.6

Table 2: Performance on RGBNT100 and MSVR310.

Comparison with State-of-the-Art Methods

Multi-spectral Person ReID. In Tab. 1, we compare our TOP-ReID with both single-spectral methods and multi-spectral methods on RGBNT201. The results indicate that single-spectral methods generally achieve lower performance compared with multi-spectral methods. It demonstrates the effectiveness of utilizing complementary information from different image spectra. Among the single-spectral methods, PCB achieves the highest performance, attaining the mAP and Rank-1 accuracy of 32.8% and 28.1%, respectively. As for the multi-spectral methods, our TOP-ReID achieves remarkable performance. Specifically, it achieves a mAP that is 24.8% higher and a Rank-1 accuracy that surpasses IEEE by 32.2%. These performance gains provide strong evidences for our TOP-ReID in tackling the challenges of multi-spectral person ReID.

Multi-spectral Vehicle ReID. As shown in Tab. 2, single-spectral methods such as OSNet (Zhou et al. 2019), AGW

Methods		M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)	
		mAP	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**
Single	HACNN	12.5	11.1	20.5	19.4	16.7	13.3	9.2	6.2	6.3	2.2	14.8	12.0
	MUDeep	19.2	16.4	20.0	17.2	18.4	14.2	13.7	11.8	11.5	6.5	12.7	8.5
	OSNet	19.8	17.3	21.0	19.0	18.7	14.6	12.3	10.9	9.4	5.4	13.0	10.2
	MLFN	20.2	18.9	21.1	19.7	17.6	11.1	13.2	12.1	8.3	3.5	13.1	9.1
	CAL	21.4	22.1	24.2	23.6	18.0	12.4	18.6	20.1	10.0	5.9	17.2	13.2
	PCB	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4
Multi	PFNet	-	-	31.9	29.8	25.5	25.8	-	-	-	-	26.4	23.4
	DENet	-	-	35.4	36.8	33.0	35.4	-	-	-	-	32.4	29.2
	TOP-ReID	**54.4**	**57.5**	**64.3**	**67.6**	**51.9**	**54.5**	**35.3**	**35.4**	**26.2**	**26.0**	**34.1**	**31.7**

Table 3: Experimental results of missing-spectral tasks on RGBNT201. "M (X)" stands for missing the X image spectra.

	Modules				RGBNT201			
	BL	**AL**	**TPM**	**CRM**	**mAP**	**R-1**	**R-5**	**R-10**
A	✓	×	×	×	55.9	54.9	70.8	77.6
B	×	✓	×	×	62.9	64.5	77.4	82.7
C	×	✓	✓	×	67.8	69.4	83.3	88.8
D	×	✓	✓	✓	72.3	76.6	84.7	89.4

Table 4: Performance comparison with different modules.

(Ye et al. 2021) and TransReID (He et al. 2021), stand out for their competitive performance. For multi-spectral methods, CCNet achieves remarkable results across both datasets. On the RGBNT100 dataset, our TOP-ReID outperforms CCNet with a 4.0% higher mAP. On the small-scale MSVR310 dataset, our TOP-ReID maintains competitive performance, showing its versatility and robustness.

Evaluation on Missing-spectral Scenarios. As shown in Tab. 3, all single-spectral methods suffer from performance degradations when image spectra are missing. Multi-spectral methods demonstrate better robustness compared with single-spectral methods. Our proposed TOP-ReID achieves remarkable performance even in the presence of missing spectra. It consistently outperforms both single-spectral and multi-spectral methods in all missing-spectral scenarios, indicating its effectiveness in handling the spectral incompleteness. In addition, compared with PFNet and DENet, our TOP-ReID is a more flexible and diverse framework to address any spectra missing.

Ablation Studies

To investigate the effect of different components, we further perform a scope of ablation studies on RGBNT201.

Effects of Key Modules. Tab. 4 illustrates the performance comparison with different modules. The Model A is the baseline which utilizes the multi-stream ViT-B/16 backbones. BL means the triplet loss and cross-entropy loss are added before the concatenation of multi-spectral features, while AL means these losses are employed after the feature concatenation. It can be observed that the AL setting shows better results. The main reason is that the fused multi-spectral features is more powerful than the simple feature concatenation. Furthermore, by integrating our TPM, the Model C yields higher performance with mAP of 67.8% and Rank-1 of 69.4%. By introducing CRM, the final model can achieve the best performance with mAP of 72.3% and Rank-1 of 76.6%. These improvements validate the effectiveness of our key modules in handling complex ReID scenarios.

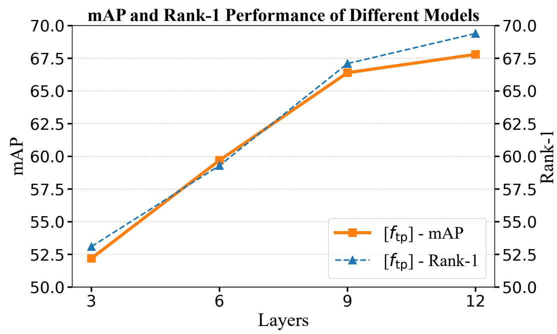


Figure 4: Performance of deploying TPM at different layers.

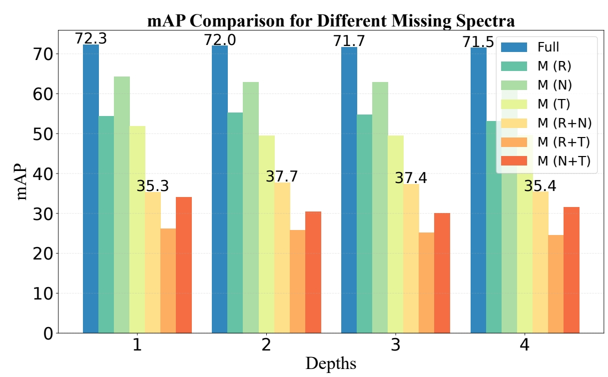


Figure 5: Effects of different depths of TransRe blocks.

TPM at Different Layers. In fact, our TPM is a plug-and-play module. We explore the effect of TPM at different layers of the ViT backbone. Fig. 4 shows the performance of TPM at different layers. We observe that as the plugged depth of TPM increases, the performance greatly improves. When deployed in the last layer, it achieves the best performance. This indicates that our TPM is more pronounced in deep layers, capturing more discriminative representations.

Effects of TransRe Blocks in CRM. The depth of TransRe blocks may impact the reconstruction ability. As illustrated in Fig. 5, the ReID performance is relatively consistent when using different depths of TransRe blocks. In

Methods	ViT-B/16				DeiT-S/16				T2T-ViT-24			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
RGB	29.0	26.2	44.5	56.1	33.3	30.6	49.5	58.0	30.1	30.3	47.2	56.6
NIR	18.7	14.0	31.1	44.9	22.7	21.4	39.4	47.6	15.8	15.3	27.0	36.8
TIR	33.4	32.3	52.4	63.3	27.1	26.3	41.3	51.4	34.0	36.2	52.0	62.0
NIR-TIR	45.9	43.4	59.9	69.4	40.6	40.9	54.3	61.0	40.9	40.7	56.3	64.2
RGB-NIR	39.0	40.2	56.6	65.7	46.7	45.0	62.6	70.0	36.3	35.2	53.8	66.3
RGB-TIR	52.6	53.8	69.0	78.2	49.3	47.8	64.1	72.8	49.9	51.7	66.7	73.8
RGB-TIR-NIR	55.9	54.9	70.8	77.6	55.1	53.3	67.3	76.2	52.2	51.3	64.1	74.3
Baseline (AL)	62.9	64.5	77.4	82.7	59.9	61.1	73.9	80.9	56.2	60.4	73.0	78.6
Baseline (AL) + TPM	67.8	69.4	83.3	88.8	63.0	63.9	78.1	83.9	58.2	60.8	74.9	81.4
Baseline (AL) + CRM + TPM	**72.3**	**76.6**	**84.7**	**89.4**	**69.0**	**73.6**	**81.8**	**84.7**	**60.0**	**61.6**	**76.2**	**82.3**

Table 5: Performance comparison of different backbones with different spectra and modules on RGBNT201.

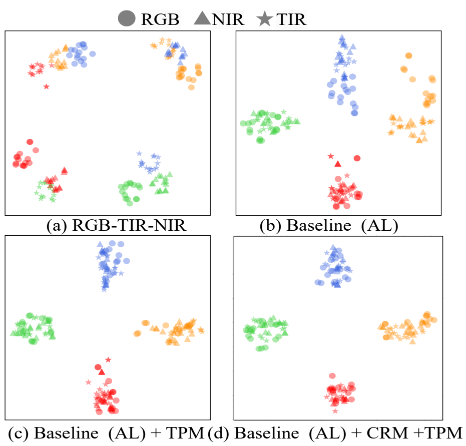


Figure 6: Comparison of feature distributions by using t-SNE. Different colors represent different identities.

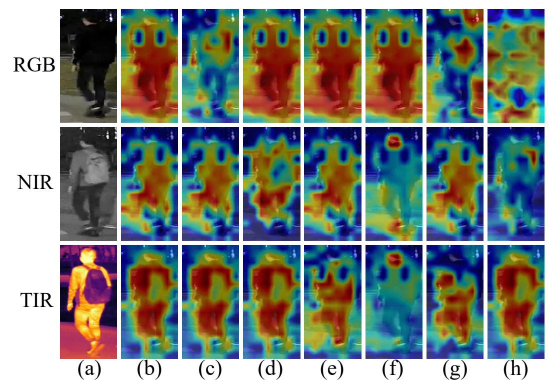


Figure 7: Discriminative attention maps. (a) Input images; (b) Full; (c) M (RGB); (d) M (NIR); (e) M (TIR); (f) M (NIR+TIR); (g) M (RGB+TIR); (h) M (RGB+NIR);

Fig. 5, we also provide the comparison results with missing-spectral cases. It can be observed that the overall performance is acceptable when only using one block. Thus, we utilize one TransRe block to reduce the computation.

Effects of Different Transformer-based Backbones. To verify the generalization of our TOP-ReID, we adopt three different Transformer-based backbones, i.e., ViT-B/16, DeiT-S/16 and T2T-ViT-24. Tab. 5 illustrates the performance comparison. As can be observed, the ViT-B/16 delivers the best results. With more image spectra, different backbones can consistently improve the performance. Our proposed TPM and CRM can improve the performance with different backbones. We believe that the performance can be further improved by using more powerful backbones.

Visualization Analysis

To clarify the learning ability, we present visual results on the feature distributions and discriminative attention maps.

Multi-spectral Feature Distributions. Fig. 6 illustrates the feature distributions of different models by using t-SNE (Van der Maaten and Hinton 2008). In Fig. 6 (a), it represents the direct concatenation of single-spectral features, where each stream is individually trained. It can be observed that the AL setting can effectively align the features of different spectra with a better ID consistence. With our TPM, the features of the same ID across different spectra are more concentrated, and the gaps between different IDs are more distinct. Furthermore, with CRM, the feature distribution becomes more compact, and the number of outliers for each ID is reduced. This visualization provides strong evidences for the effectiveness of our proposed methods.

Discriminative Attention Maps. As shown in Fig. 7, we utilize Grad-CAM (Selvaraju et al. 2017) to visualize the discriminative attention maps with different image spectra. Obviously, there are discriminative differences between different image spectra. Our model is powerful and can highlight discriminative regions when missing image spectra.

Conclusion

In this work, we propose a novel feature learning framework based on token permutations for multi-spectral object ReID. Our approach incorporates a Token Permutation Module (TPM) for spatial feature alignment and a Complementary Reconstruction Module (CRM) for reducing the distribution gap across different image spectra. Through the dynamic cooperation between TPM and CRM, it can handle the missing-spectral problem, which is more flexible than previous methods. Extensive experiments on three benchmarks clearly demonstrate the effectiveness of our methods.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2018AAA0102001), National Natural Science Foundation of China (No. 62101092), Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202102) and Fundamental Research Funds for the Central Universities (No. DUT22QN228).

References

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

Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. In *CVPR*, 2109–2118.

Chen, G.; Zhang, T.; Lu, J.; and Zhou, J. 2019. Deep meta metric learning. In *ICCV*, 9547–9556.

Chen, Y.; Xia, S.; Zhao, J.; Zhou, Y.; Niu, Q.; Yao, R.; Zhu, D.; and Liu, D. 2022. ResT-ReID: Transformer block-based residual learning for person re-identification. *PRL*, 157: 90–96.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guo, J.; Zhang, X.; Liu, Z.; and Wang, Y. 2022. Generative and attentive fusion for multi-spectral vehicle re-identification. In *ICSP*, 1565–1572.

He, Q.; Lu, Z.; Wang, Z.; and Hu, H. 2023. Graph-Based Progressive Fusion Network for Multi-Modality Vehicle Re-Identification. *TITS*, 1–17.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *ICCV*, 15013–15022.

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

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020a. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, volume 34, 4610–4617.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020b. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, volume 34, 11345–11353.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*, 2285–2294.

Liu, H.; Chai, Y.; Tan, X.; Li, D.; and Zhou, X. 2021a. Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification. *SPL*, 28: 653–657.

Liu, X.; Yu, C.; Zhang, P.; and Lu, H. 2023. Deeply Coupled Convolution–Transformer With Spatial–Temporal Complementary Learning for Video-Based Person Re-Identification. *TNNLS*.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021b. Watching You: Global-Guided Reciprocal Learning for Video-Based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13334–13343.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 1487–1495.

Pan, W.; Huang, L.; Liang, J.; Hong, L.; and Zhu, J. 2023. Progressively Hybrid Transformer for Multi-Modal Vehicle Re-Identification. *Sensors*, 23(9): 4206.

Pan, W.; Wu, H.; Zhu, J.; Zeng, H.; and Zhu, X. 2022. H-vit: Hybrid vision transformer for multi-modal vehicle re-identification. In *CICAI*, 255–267. Springer.

Qian, X.; Fu, Y.; Jiang, Y.-G.; Xiang, T.; and Xue, X. 2017. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 5399–5408.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, 1025–1034.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 6398–6407.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 480–496.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 274–282.

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022a. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, 7297–7307.

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022b. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, volume 36, 2633–2641.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 44(6): 2872–2893.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 558–567.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 516–525.

Zhang, Y.; and Wang, H. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *CVPR*, 2153–2162.

Zhao, J.; Zhao, Y.; Li, J.; Yan, K.; and Tian, Y. 2021. Heterogeneous relational complement for vehicle re-identification. In *ICCV*, 205–214.

Zheng, A.; He, Z.; Wang, Z.; Li, C.; and Tang, J. 2023. Dynamic Enhancement Network for Partial Multi-modality Person Re-identification. *arXiv preprint arXiv:2305.15762*.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust multi-modality person re-identification. In *AAAI*, volume 35, 3529–3537.

Zheng, A.; Zhu, X.; Ma, Z.; Li, C.; Tang, J.; and Ma, J. 2022. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark. *arXiv preprint arXiv:2208.00632*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, volume 34, 13001–13008.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-scale feature learning for person re-identification. In *ICCV*, 3702–3712.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2223–2232.

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