# PEFT-BoA: Parameter-Efficient Fine-Tuning with Bag-of-Adapters for Multi-Modal Object Re-identification

**Hongchao Li[1], Guangxing Liu[1], Xixi Wang[2*], Baihe Liang[1], YongLong Luo[1]**
[1]School of Computer and Information Science, Anhui Normal University
Wuhu, Anhui Province 241002, China
[2]School of Computer Science and Technology, Anhui University
Hefei, Anhui Province 230601, China
{lhc950304, sissiw0409, bh.liang}@foxmail.com, ffunly@163.com, ylluo@ustc.edu.cn

## Abstract

Multi-modal object Re-identification (ReID) aims to retrieve individuals by leveraging complementary information from different modalities. Recent CLIP-based approaches show promising results, but they usually employ prompt-based or hybrid prompt-adapter tuning and still face the problems of heterogeneous domain gap, fine-grained identity discrimination and noise instance interference. To address these problems, we introduce a novel Parameter-Efficient Fine-Tuning framework with Bag-of-Adapters (PEFT-BoA) based on the pre-trained CLIP's vision encoder for multi-modal object ReID. Specifically, we first propose a Domain-specific Patch Adapter (DPA) designed to bridge the visual feature gap between pre-trained and fine-tuned models at the local patch level. Meanwhile, we propose a Task-specific Class Adapter (TCA) enhance the fine-grained identity discrimination ability by optimizing global class token. Finally, we propose an Instance-specific Fusion Adapter (IFA) dynamically selects and combines only the most useful features across different modalities for each instance. Our PEFT-BoA achieves the better performance on multi-modal object re-identification benchmarks, while maintaining fewer trainable parameters (6.62M) and a higher training throughput (246.2fps).

**Code** — https://github.com/fffunly/PEFT-BoA

## Introduction

Multi-modal object Re-identification (ReID) is a critical task in computer vision, with applications ranging from surveillance to autonomous systems. The goal is to retrieve objects across different modalities, such as RGB, Near-Infrared (NIR), and Thermal-Infrared (TIR), by learning discriminative feature representations. However, this task faces significant challenges, including the significant domain divergence of the input modalities from the pre-training data, limited annotated multi-modal data, and the need for fine-grained discrimination. Traditional approaches often struggle to generalize efficiently across diverse modalities while maintaining computational feasibility, creating a bottleneck for real-world deployment.

Earlier methods (Li et al. 2020; Crawford et al. 2023) primarily rely on Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) to achieve the competitive performances. However, they typically train from scratch or fully fine-tuning on task-specific datasets. Although the results are satisfactory, the full fine-tuning paradigm requires substantial computational resources and large-scale labeled data, which makes it inefficient for multi-modal adaptation. Recent advances in large-scale pre-trained models, such as CLIP (Radford et al. 2021; Li, Sun, and Li 2023), offer a promising alternative. CLIP's capabilities in cross-modal understanding and representation learning provides a robust foundation for multi-modal object ReID. However, directly applying CLIP without adaptation still faces challenge, particularly in domain shift and task-specific optimization.

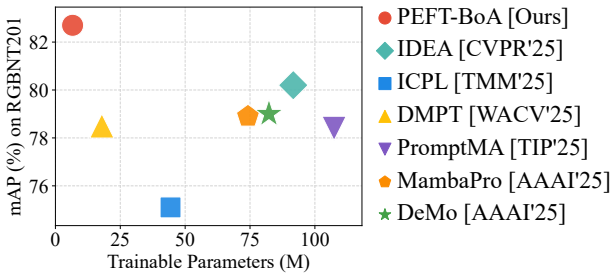


Figure 1: Comparison of trainable parameters and validation accuracy between our PEFT-BoA and some other state-of-the-art CLIP-based methods on the RGBNT201 dataset.

Recent Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged as a practical way to adapt CLIP for downstream tasks with fewer trainable parameters. As shown in Fig. 1, including adapter-based tuning (Wang et al. 2025a) and prompt-based tuning (Lin et al. 2025), preserve the model's pre-trained knowledge while reducing computational overhead. Several studies (Wang et al. 2025a,c) have explored CLIP-based approaches for multi-modal object ReID and demonstrate the promising results. For example, (Wang et al. 2025a) develop a MambaPro framework based on mamba aggregation and synergistic prompt tuning for multi-modal object ReID. (Lin et al. 2025) propose a Decoupled Modality-aware Prompt Tuning (DMPT) framework, which mainly employs a prompt inverse bind strategy to mine the complementary cues among different modalities for multi-modal object ReID. (Zhang et al. 2025) in-

*Corresponding author.

troduce a Prompt-based Modality Alignment (PromptMA) framework, which capture the modal complementary feature across different modality paths to reduce the modal distribution gap and thus achieve more effective multi-modal fusion. However, these methods still suffer from several main limitations: First, the prevailing PEFT methods impose an isomorphic adaptation on all tokens, disregarding the inherent functional heterogeneity between the class and patch tokens. Second, the entangled optimization of cross-modal alignment and fine-grained discrimination in a single prompt space may lead to suboptimal ReID performance. Third, they typically enforce full-instance interaction across modalities (e.g., multi-head self-attention mechanism), which not only incurs unnecessary computational overhead but also risks introducing noise from irrelevant instances during feature fusion.

To address these limitations, we propose a novel Parameter-Efficient Fine-Tuning framework with Bag-of-Adapters for multi-modal object ReID task, termed as PEFT-BoA. Specifically, we first design a Domain-specific Patch Adapter (DPA), which learns a unique spatial transformation for each input modality (e.g., NIR, TIR) to bridge the vast domain gap to CLIP's pre-trained feature space. Then, we introduce a Task-specific Class Adapter (TCA) to decouple the optimization objectives by separately enhancing fine-grained discriminative ability while preserving CLIP's general representation capability. Finally, we develop an Instance-specific Fusion Adapter (IFA) which incorporates a SwiGLU-based gating mechanism to dynamically select relevant cross-modal features. This aims to mitigate noise from irrelevant correlations and overcome computational waste from image redundancy (e.g., background, similar patches), thus maintaining high efficiency, as inspired by mPLUG-Owl2 (Ye et al. 2024). The whole network architecture is trained in an end-to-end way. Through extensive experiments, we demonstrate that the proposed PEFT-BoA achieves better performance with significantly fewer parameters than the existing methods.

Overall, our key contributions can be summarized as:

- We design a novel Parameter-Efficient Fine-Tuning framework with Bag-of-Adapters for multi-modal object ReID. To the best of our knowledge, this is the first successful adaption of CLIP model for multi-modal object ReID using purely adapter-based fine-tuning.
- We propose a novel Domain-specific Patch Adapter (DPA) that aligns local patch-level features across different modalities, a Task-specific Class Adapter (TCA) that optimizes global class token for fine-grained identity discrimination and an Instance-specific Fusion Adapter (IFA) that enables adaptive fusion of complementary information from different modalities.
- Our method achieves the competitive performance on four multi-modal object ReID benchmarks while maintaining significantly lower parameters ($<$10% of full fine-tuning), demonstrating an optimal balance between performance and efficiency.

## Related Work

Multi-modal object Re-identification (ReID) aims to accurately match objects across multiple modalities. Recent advances can be broadly categorized into full fine-tuning and partial fine-tuning paradigms.

**Full Fine-Tuning:** This paradigm optimize all parameters of pre-trained models on object ReID datasets, offering strong task-specific adaptation but requiring substantial computational resources. It encompasses CNN-, ViT-, and CLIP-based methods, each adapting the entire backbone to align multi-modal features (e.g., RGB, NIR, or TIR). Early works (Zheng et al. 2021; Wang et al. 2022; Zheng et al. 2025) primarily rely on CNNs (e.g., ResNet (He et al. 2016)) fine-tuned end-to-end for multi-modal object ReID, often integrating modality-specific branches or fusion modules. With the rise of vision transformers, ViT-based methods (Wang et al. 2024a; Zhang et al. 2024; Wang et al. 2024b) extend this paradigm, leveraging full fine-tuning of self-attention layers to model cross-modal interactions at the cost of quadratic computational complexity. CLIP-based methods (Feng et al. 2025; Wang et al. 2025b; Wan et al. 2025) have recently demonstrated promise by fully fine-tuning both its visual encoders for multi-modal object ReID. For example, (Zhang et al. 2025) fully fine-tune the model with prompt-based alignment to bridge modality gaps. (Wang et al. 2025b) dynamically balance multi-modal features via mixture-of-experts. Notably, these methods often overlook CLIP's inherent robustness to distribution shifts, as full fine-tuning may distort its pre-trained cross-modal alignment.

**Partial Fine-Tuning:** This paradigm aims to preserve most pre-trained knowledge while adapting only task-specific components, offering better efficiency and generalization. Unlike full fine-tuning, to the best of our knowledge, it mainly relies on CLIP-based framework in multi-modal object ReID. The reason for this distinction may stem from the robust cross-modal alignment that CLIP learns from large-scale vision-language data. However, aggressive fine-tuning can easily degrade its performance, thereby motivating the use of parameter-efficient adaptation strategies. For instance, (Li et al. 2025) employ learnable text prompts and multi-spectral adapters for fine-grained multi-modal fusion. (Wang et al. 2025a) combine feed-forward adapters with Mamba aggregation for long-sequence multi-modal processing. (Lin et al. 2025) decouple visual prompts into modality-specific/independent components to enhance complementary information integration.

While existing partial tuning methods improve efficiency, their reliance on prompt mechanisms and hybrid prompt-adapter strategies constrain performance in multi-modal object ReID. Our work advances this paradigm through a pure adapter-based framework.

## Methodology

In this paper, we propose a parameter-efficient fine-tuning framework for multi-modal object ReID, introducing three novel adapters to enhance model adaptability while minimizing computational overhead. Our approach builds upon

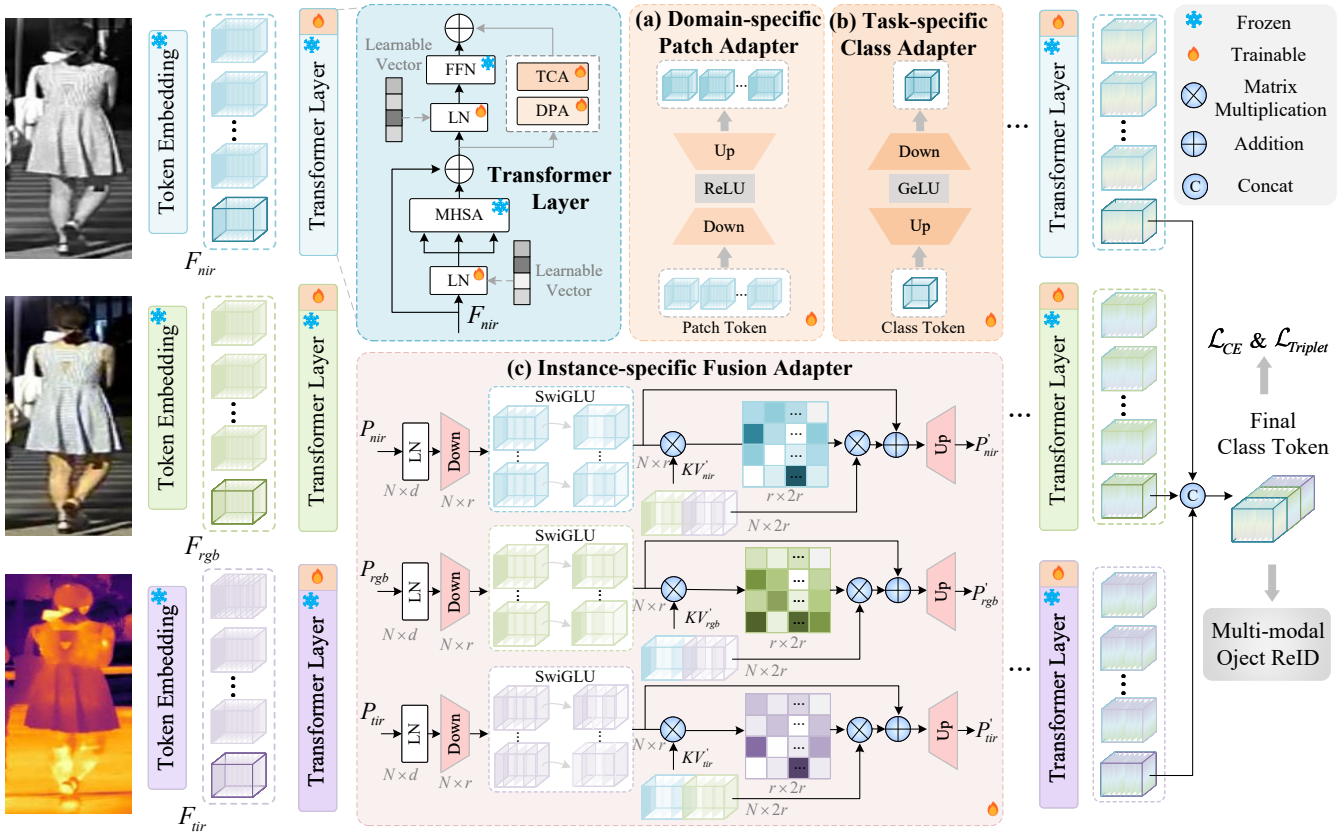


Figure 2: The overall framework of PEFT-BoA for multi-modal object ReID.

a fully frozen CLIP ViT-Base model, integrating lightweight yet effective adapters to handle domain shifts, task-specific nuances, and instance-level feature fusion.

## Baseline

We employ a CLIP ViT-Base backbone architecture, comprising standard transformer components: Multi-Head Self-Attention (MHSA) layers, Feed-Forward Networks (FFN), and Layer Normalization (LN) modules. In our baseline, no parameters within these layers are updated, ensuring maximal preservation of the pre-trained knowledge while minimizing computational overhead. During the forward pass, the three input modalities (RGB, NIR, and TIR) are processed independently through the shared frozen backbone. Within each transformer block, the operations for an input feature $F_{in}$ can be formally described as:

$$\begin{aligned} F_{mhsa} &= \text{MHSA}(\text{LN}(F_{in})) + F_{in}, \\ F_{out} &= \text{FFN}(\text{LN}(F_{mhsa})) + F_{mhsa}, \end{aligned} \tag{1}$$

where LN applies static normalization with frozen scale and shift parameters, MHSA uses fixed attention weights, and FFN retains its pre-trained projections. The final feature representation is constructed by extracting the class tokens $c_{rgb}, c_{nir}, c_{tir}$ from each modality, normalizing them via LN layers, and concatenating:

$$c_{\text{final}} = \text{Concat}(\text{LN}(c_{rgb}); \text{LN}(c_{nir}); \text{LN}(c_{tir})), \tag{2}$$

where $\text{Concat}(\cdot)$ represents the concatenation operation. To maintain simplicity and isolate the impact of the frozen backbone, we optimize the model using only a basic combination of label-smoothed cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and triplet loss ($\mathcal{L}_{\text{Triplet}}$). The total loss is defined as:

$$\mathcal{L}_{\text{Total}} = \lambda \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Triplet}}, \tag{3}$$

where $\lambda$ is a balancing hyperparameter. While this frozen baseline enables parameter-efficient inference, its rigid architecture inherently limits adaptability to cross-modal domain shifts, fine-grained discrimination, and selective feature fusion. To overcome these limitations without compromising efficiency, we introduce three novel adapters.

## Domain-specific Patch Adapter (DPA)

To address the fundamental challenge of aligning CLIP's RGB-trained representations with diverse modalities (RGB/NIR/TIR) in ReID tasks. We design a Domain-specific Patch Adapter (DPA) to operate exclusively on patch tokens, as illustrated in Fig. 2 (a). The DPA architecture implements parallel adapter modules alongside each transformer block's FFN layer, utilizing ReLU activation to achieve two key objectives. First, it functions as a non-linear filter that selectively enhances discriminative cross-modal features while suppressing noisy activations. Second, it operates as a domain alignment operator that projects different modalities

into a unified feature space by leveraging ReLU's inherent sparsity-inducing property. The detailed operations are as follows:

$$[c_{mhsa}; P_{mhsa}] = \text{Split}(F_{mhsa}), \tag{4}$$

where $F_{mhsa} \in \mathbb{R}^{(N+1)\times d}$ represents the MHSA output containing both local patch features and global context from class token interactions. $c_{mhsa} \in \mathbb{R}^{1\times d}$ and $P_{mhsa} \in \mathbb{R}^{N\times d}$ are the class token and patch tokens, respectively. Split $(\cdot)$ is an operation that separates the feature tokens. Notably, the class token in $F_{mhsa}$ incorporates the learnable vector as shown in Eq. (8). We process these attention-enhanced patch tokens $P_{mhsa}$ through our Domain-specific Patch Adapter:

$$\text{DPA}(P_{mhsa}) = \phi_u(\text{ReLU}(\phi_d(P_{mhsa}))), \tag{5}$$

where $\phi_d \in \mathbb{R}^{d\times r}$ and $\phi_u \in \mathbb{R}^{r\times d}$ are our learned projection matrices. We construct the DPA output by combining three components:

$$P_{out} = \text{FFN}(\text{LN}(P_{mhsa})) + \text{DPA}(P_{mhsa}) + P_{mhsa}, \tag{6}$$

where $P_{out}$ preserves the original CLIP representations while adding our modality-shared adaptations. The ReLU activation in DPA ensures selective enhancement of discriminative cross-modal features while suppressing noise, and the bottleneck design ($r \ll d$) maintains parameter efficiency. Our DPA module preserves the vast majority of the pre-trained model's parameters in their frozen state, only introducing comparatively small trainable components.

## Task-specific Class Adapter (TCA)

To bridge the task gap between CLIP's pretraining and fine-grained ReID objectives, we propose the Task-specific Class Adapter (TCA). As shown in Fig. 2 (b), TCA consists of two complementary components: (1) learnable adaptation vectors for class-aware attention fine-tuning and (2) a shared adapter for class token fine-tuning.

**Learnable adaptation vectors for class-aware attention fine-tuning.** Given the transformer block input $F_{in} \in \mathbb{R}^{(N+1)\times d}$, we first isolate the class token:

$$[c_{in}; P_{in}] = \text{Split}(F_{in}), \tag{7}$$

where $c_{in} \in \mathbb{R}^{1\times d}$ represents the feature vector of the class token. Building on insights from PASTE (Yang et al. 2023), we adapt their parameter-efficient token tuning strategy for vision transformers. While PASTE introduces learnable vectors for NLP token embeddings in Post-LN architectures, we modify this approach for our Pre-LN visual backbone to enhance class-aware attention. Specifically, a class token modulation vector $v_{attn1}$ injected before the MHSA module:

$$\begin{aligned} c_{LN} &= \text{LN}(c_{in}) + v_{attn1}, \\ \overline{F}_{mhsa} &= \text{MHSA}(\text{Concat}(c_{LN}, LN(P_{in}))) + F_{in}, \end{aligned} \tag{8}$$

where $v_{attn1} \in \mathbb{R}^{1\times d}$ provides the tokens with an additional trainable bias, enabling them to capture more nuanced and salient features during attention score computation. To further enrich representations after the attention, we apply a similar strategy to the LN layer:

$$\begin{aligned} [\overline{c}_{mhsa}; \overline{P}_{mhsa}] &= \text{Split}(\overline{F}_{mhsa}), \\ \overline{c}_{LN} &= \text{LN}(\overline{c}_{mhsa}) + v_{attn2}, \end{aligned} \tag{9}$$

where $v_{attn2} \in \mathbb{R}^{1\times d}$ further refines the attention output before processing them in the subsequent FFN layer. Together, $v_{attn1}$ and $v_{attn2}$ create a dual modulation mechanism that progressively shapes both attention patterns and feature representations to achieve more effective class-aware discrimination, thereby systematically enhancing the model's fine-grained semantic understanding capacity.

**Shared adapter for class token fine-tuning.** Recognizing that the class token represents only a simple vector, we avoid layer-specific adapters to prevent parameter inflation and overfitting. Instead, we propose a unified adapter that operates in parallel with each FFN layer while sharing parameters across all transformer blocks. The shared-weight architecture ensures consistent class token adaptation throughout the network, contrasting with DPA's patch-focused approach yet complementing it through specialized token processing. Formally, let $\overline{c}_{mhsa}$ denote the class token after the modulated MHSA output. The class token is then processed separately through our adapter as follows:

$$\text{TCA}(\overline{c}_{mhsa}) = \phi_{\text{d}}(\text{GELU}(\text{Dropout}(\phi_{\text{u}}(\overline{c}_{mhsa})))), \tag{10}$$

where projection matrices $\phi_{\text{d}}$ and $\phi_{\text{u}}$ are shared across all transformer layers and input modalities (RGB/NIR/TIR), ensuring parameter efficiency while maintaining consistent adaptation. GELU activation provides smoother gradients than ReLU while preserving sparse activations, crucial for preventing dead neurons in the projection space. Dropout regularizes the modality-specific projections by randomly masking activations during training, preventing co-adaptation between modalities. The TCA output for the class token is computed as:

$$c_{out} = \text{FFN}(\overline{c}_{LN}) + \text{TCA}(\overline{c}_{mhsa}) + \overline{c}_{mhsa}, \tag{11}$$

where $c_{out}$ enhances global semantics via shared cross-layer and cross-modal projections, forming a complementary 'class-patch' adaptation with the DPA output.

## Instance-specific Fusion Adapter (IFA)

Multi-modal fusion is crucial for multi-modal object ReID, but existing fusion modules often introduce substantial computational overhead. To align with our lightweight, plug-and-play framework, we propose an Instance-specific Fusion Adapter (IFA), inspired by recent advances like the BAT adapter (Cao et al. 2024) and Mamba adapter (Shi et al. 2025) in multi-modal object tracking. Fig. 2 (c) illustrates the detailed architecture of IFA.

Given patch tokens $P_{out} \in \mathbb{R}^{N\times d}$ from the preceding Transformer block, let $P_m$ denote modality-specific patch tokens (RGB/NIR/TIR). The RGB feature enhancement process begins with layer normalization and SwiGLU (Gated Linear Unit with Swish activation) projections:

$$\begin{aligned} \overline{P}_{\text{rgb}} &= \text{LN}(P_{\text{rgb}}), \\ Q_{\text{rgb}} &= \text{SiLU}(\phi_{q,g}(\overline{P}_{\text{rgb}})) \odot (\phi_{q,d}(\overline{P}_{\text{rgb}})), \\ KV_{\text{rgb}} &= \text{SiLU}(\phi_{kv,g}(\overline{P}_{\text{rgb}})) \odot (\phi_{kv,d}(\overline{P}_{\text{rgb}})), \end{aligned} \tag{12}$$

where $\odot$ denotes the Hadamard product. The weights $\phi_{q,g}$ and $\phi_{q,d}$ parameterize the RGB-specific gating and projec-

tion operations (similarly for $\phi_{kv,g}$, $\phi_{kv,d}$). The SiLU activation $\text{SiLU}(x) = x\sigma(x)$ combines adaptive feature selection ($\sigma(x)$) with gradient preservation ($x$). For a specific instance, other modalities (NIR/TIR) undergo identical processing to generate their respective key-value pairs $KV_{\text{nir}}, KV_{\text{tir}}$. For cross-modal interaction, we concatenate the complementary modalities' key-value pairs as context, then compute the enhanced RGB features through a residual attention mechanism:

$$\begin{aligned} KV'_{\text{rgb}} &= \text{Concat}(KV_{\text{nir}}, KV_{\text{tir}}), \\ Q'_{\text{rgb}} &= Q_{\text{rgb}} + \alpha \cdot \text{Attention}(Q_{\text{rgb}}, KV'_{\text{rgb}}, KV'_{\text{rgb}}), \end{aligned} \tag{13}$$

where the scaling factor $\alpha$ controls the cross-modal fusion intensity. The enhanced query representation $Q'_{\text{rgb}}$ undergoes dimension restoration through an up-projection layer $\phi_{\text{u}}$, followed by residual fusion with the original patch tokens:

$$\begin{aligned} P'_{\text{rgb}} &= \phi_{\text{u}}(Q'_{\text{rgb}}) + P_{\text{rgb}}, \\ F'_{\text{rgb}} &= \text{Concat}(c_{\text{rgb}}, P'_{\text{rgb}}). \end{aligned} \tag{14}$$

This processed feature $F'_{\text{rgb}}$ serves as input to the next Transformer block. This symmetric processing pipeline is identically applied to other modalities (NIR/TIR), creating a balanced architecture that enables mutual feature enhancement through cross-modal attention. The design achieves effective feature fusion across modalities, with residual connections ensuring stable gradient flow during training. Remarkably, this module maintains computational efficiency, introducing minimal parameter overhead while significantly improving cross-modal feature consistency.

# Experiments

## Experimental Setup

**Datasets and Evaluation Protocols.** To comprehensively evaluate the effectiveness of our proposed method, we conduct extensive experiments on four widely-used multi-modal object ReID benchmarks, such as RGBN201 (Zheng et al. 2021), Market-MM (Wang et al. 2022), RGBNT100 (Li et al. 2020) and MSVR310 (Zheng et al. 2023). Concretely, **RGBN201** is the first multi-modal person ReID dataset, which is captured on campus by using four non-overlapping cameras. It comprises 201 identities with 4787 aligned image triples of three modalities, split into 141 identities for training, 30 for validation and 30 for testing. **Market-MM** is a synthetic multi-modal person Re-ID dataset that extends Market-1501 (Zheng et al. 2015) by generating corresponding thermal, near-infrared, and night-time images for the original RGB data. **RGBNT100** focuses on multi-modal vehicle ReID across complex outdoor surveillance scenes involving 8 cameras. It contains 17250 image triples of 100 vehicles, which is split into 8675 image triples of 50 vehicles for training and 8575 image triples of other 50 vehicles for testing. **MSVR310** is a smaller but highly challenging multi-modal vehicle ReID dataset, which contains 2087 images of 310 vehicles collected by four cameras. It randomly select 1032 image triples of 155 vehicles for training and 1055 image triples of 155 vehicles for testing. For a fair comparison, we conduct evaluations on the above-mentioned four datasets using the standard metrics, i.e., Cumulative Matching Characteristic (CMC) curve at Rank-1 (R-1), Rank-5 (R-5), Rank-10 (R-10) and mean Average Precision (mAP). The best and second best results are marked in **bold** and underline, respectively.

**Implementation Details** We implement our framework using the pre-trained CLIP's image encoder (Radford et al. 2021) as the frozen backbone. All original model parameters remain fixed. Optimization is performed exclusively on our proposed modules. For parameter initialization, all learnable vectors within these adapters are initialized to zero, while other components are initialized following the Houlsby method (Houlsby et al. 2019), employing a truncated normal distribution. We resize the resolution of input images to $256 \times 128$ for person datasets, while $128 \times 256$ for vehicle datasets. We apply the standard data augmentation techniques (e.g., random horizontal flipping, cropping, and random erasing (Zhong et al. 2020)) during training to improve model generalization. For RGBNT201 and MSVR310 (smaller datasets), we randomly select 16 identities per batch and sample 4 images for each identity, i.e., batch size is set to 64. For Market-MM and RGBNT100 (larger datasets), we randomly choose 8 identities per batch with 16 images each, i.e., batch size is set to 128. We employ the AdamW optimizer with an initial learning rate of $3 \times 10^{-4}$ and use the warmup strategy with a cosine decay for learning rate scheduling to fine-tuning model. The total number of training epochs is set to 120 for all datasets. All our experiments are conducted using the PyTorch on a single NVIDIA RTX 4090 24G GPU.

## Performance Comparison

**Multi-modal Person ReID.** To evaluate the effectiveness of the proposed PEFT-BoA, we conduct a comprehensive comparison with some other state-of-the-art approaches on two widely used multi-modal person ReID benchmarks (i.e., RGBNT201 and Market-MM). The experimental results are presented in Table 1. We can observe that our PEFT-BoA achieves the highest performance on both datasets when compared with CNN-based and ViT-based methods using full fine-tuning. Specifically, while CNN-based methods such as IEEE (Wang et al. 2022) attain moderate performance, and ViT-based methods like EDITOR (Zhang et al. 2024) demonstrate improvements, our method outperforms them substantially with an mAP of 82.7% and Rank-1 accuracy of 86.1% on RGBNT201, alongside 86.7% mAP and 94.4% Rank-1 on Market-MM. For the sake of fairness, we also compare our method with those based on CLIP. We can observe that our PEFT-BoA outperforms these methods by approximately 1 to 4 percentage points in mAP and Rank-1 metrics. For example, on RGBNT201 dataset, compared with DMPT method (Lin et al. 2025) based on prompt-tuning, our PEFT-BoA exceed it by +4.2% in mAP and +4.8% in the Rank-1 metric. Compared with MambaPro (Wang et al. 2025a) method based on the hybrid fine-tuning of adapter and prompt, our PEFT-BoA outperforms it by +3.8% in mAP and +2.7% in the Rank-1 metric. With only 6.62M parameters, our approach requires $<3\%$ of the parameters used in TOP-ReID (324.53M) or UniCat

| Methods | Backbone | Para. (M) | RGBNT201 | | | | Market-MM | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP | R-1 | R-5 | R-10 | mAP | R-1 | R-5 | R-10 |
| HAMNet (Li et al. 2020) | CNN | 78.00 | 27.7 | 26.3 | 41.5 | 51.7 | 60.0 | 82.8 | - | - |
| PFNet (Zheng et al. 2021) | CNN | - | 38.5 | 38.9 | 52.0 | 58.4 | 60.9 | 83.6 | 92.8 | 95.5 |
| IEEE (Wang et al. 2022) | CNN | 109.22 | 46.4 | 47.1 | 58.5 | 64.2 | 64.3 | 83.9 | 93.0 | 95.7 |
| UniCat (Crawford et al. 2023) | ViT | 259.02 | 57.0 | 55.7 | - | - | - | - | - | - |
| TOP-ReID (Wang et al. 2024a) | ViT | 324.53 | 72.3 | 76.6 | 84.7 | 89.4 | 82.0 | 92.4 | 97.6 | 98.6 |
| EDITOR (Zhang et al. 2024) | ViT | 118.55 | 66.5 | 68.3 | 81.1 | 88.2 | 77.4 | 90.8 | 96.8 | 98.3 |
| PromptMA (Zhang et al. 2025) | CLIP | 107.4 | 78.4 | 80.9 | 87.0 | 88.9 | 83.6 | 93.3 | - | - |
| DMPT (Lin et al. 2025) | CLIP | 17.90 | 78.5 | 81.3 | 90.4 | 93.5 | 82.7 | 92.0 | 97.1 | 98.5 |
| MambaPro (Wang et al. 2025a) | CLIP | 74.20 | 78.9 | 83.4 | 89.8 | 91.9 | 84.1 | 92.8 | 97.7 | 98.7 |
| DeMo (Wang et al. 2025b) | CLIP | 98.79 | 79.0 | 82.3 | 88.8 | 92.0 | 83.6 | 93.1 | 97.5 | 98.7 |
| ICPL-ReID (Li et al. 2025) | CLIP | 44.36 | 75.1 | 77.4 | 84.2 | 87.9 | 85.1 | **94.7** | **98.4** | **99.1** |
| IDEA (Wang et al. 2025c) | CLIP | 91.67 | 80.2 | 82.1 | 90.0 | 93.3 | - | - | - | - |
| **PEFT-BoA (Ours)** | CLIP | 6.62 | **82.7** | **86.1** | **92.3** | **94.7** | **86.7** | 94.4 | 98.2 | **99.1** |

Table 1: Comparison with the state-of-the-art methods on RGBNT201 and Market-MM.

| Methods | RGBNT100 | | MSVR310 | |
|---|---|---|---|---|
| | mAP | R-1 | mAP | R-1 |
| PFNet (Zheng et al. 2021) | 68.1 | 94.1 | 23.5 | 37.4 |
| HAMNet (Li et al. 2020) | 74.5 | 93.3 | 27.1 | 42.3 |
| TOP-ReID (Wang et al. 2024a) | 81.2 | 96.4 | 35.9 | 44.6 |
| EDITOR (Zhang et al. 2024) | 82.1 | 96.4 | 39.0 | 49.3 |
| RSCNet (Yu et al. 2025) | 82.3 | 96.6 | 39.5 | 49.6 |
| DMPT (Lin et al. 2025) | 81.7 | 94.1 | 36.6 | 52.1 |
| MambaPro (Wang et al. 2025a) | 83.9 | 94.7 | 47.0 | 56.5 |
| DeMo (Wang et al. 2025b) | 86.2 | 97.6 | 49.2 | 59.8 |
| IDEA (Wang et al. 2025c) | **87.2** | 96.5 | 47.0 | 62.4 |
| PromptMA (Zhang et al. 2025) | 85.3 | 97.4 | 55.2 | 64.5 |
| ICPL-ReID (Li et al. 2025) | 87.0 | **98.6** | 56.9 | **77.7** |
| **PEFT-BoA (Ours)** | 85.1 | 97.8 | **57.4** | 74.5 |

Table 2: Performance on RGBNT100 and MSVR310.

| # | Components | | | RGBNT201 | | | |
|---|---|---|---|---|---|---|---|
| | DPA | TCA | IFA | mAP | R-1 | R-5 | R-10 |
| 1 | × | × | × | 2.4 | 0.8 | 2.9 | 4.4 |
| 2 | ✓ | × | × | 77.0 | 78.2 | 88.2 | 92.1 |
| 3 | × | ✓ | × | 63.5 | 66.0 | 79.5 | 88.5 |
| 4 | × | × | ✓ | 62.9 | 64.7 | 76.1 | 81.7 |
| 5 | ✓ | ✓ | × | 81.1 | 83.7 | 90.3 | 93.7 |
| 6 | ✓ | × | ✓ | 76.8 | 79.4 | 87.8 | 90.8 |
| 7 | × | ✓ | ✓ | 72.9 | 77.9 | 87.3 | 91.1 |
| 8 | ✓ | ✓ | ✓ | 82.7 | 86.1 | 92.3 | 94.7 |

Table 3: Ablation study on RGBNT201.

(259.02M). In summary, the experimental results show that our PEFT-BoA is not only highly effective in multi-modal person ReID but also offers an advantageous trade-off between model size and retrieval performance.

**Multi-modal Vehicle ReID.** As depicted in Table 2, we further validate the effectiveness of our PEFT-BoA method on RGBNT100 and MSVR310 datasets. Specifically, our method attains competitive performance, achieving an mAP of 85.1% and Rank-1 accuracy of 97.8% on RGBNT100. On MSVR310, our method achieves a strong mAP of 57.4% and Rank-1 accuracy of 74.5%, demonstrating robust cross-modal feature learning in this challenging setting. Compared with MambaPro (Wang et al. 2025a), which employs a hybrid fine-tuning strategy combining adapter and prompt, our PEFT-BoA achieves improvements of +1.2% in mAP and +3.1% in Rank-1 metric. This further verifies the generalizability and practicality of our PEFT-BoA across multi-modal object ReID tasks beyond person datasets.

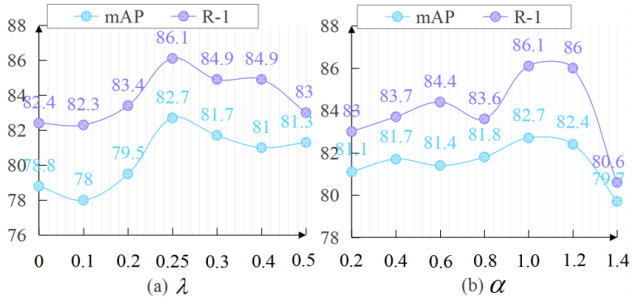


Figure 3: Parameter analysis results on RGBNT201.

## Ablation Study

**Different Components Analysis.** To analyze the effectiveness of three key components in our PEFT-BoA framework, we conduct an ablation study on RGBNT201 dataset. Table 3 reports the comparison results where each row selectively includes or excludes these components in our PEFT-BoA model. A checkmark indicates the inclusion of the corresponding module, while a cross denotes its exclusion. '#1' serves as the baseline model, disabling all three components and corresponding to the fully frozen CLIP backbone with-

| Method | Parameter↓ (M) | GFLOPs↓ | Training Time↓ (s) | Training Memory Cost↓ (M) | Throughput↑ (fps) |
|---|---|---|---|---|---|
| DeMo | 98.79 | 34.2 | 15.7 | 12124.8 | 216.3 |
| MambaPro | 74.20 | 51.3 | 21.8 | 16195.8 | 167.0 |
| IDEA | 91.67 | 43.7 | 18.2 | 16865.1 | 186.7 |
| PromptMA | 107.40 | 35.9 | 16.9 | 15956.3 | 216.5 |
| ICPL-ReID | 44.36 | 39.8 | 21.7 | 13787.4 | 168.2 |
| Ours | **6.62** | **31.5** | **14.8** | **9002.2** | **246.2** |

Table 4: Complexity comparison between our PEFT-BoA and existing CLIP-based methods.

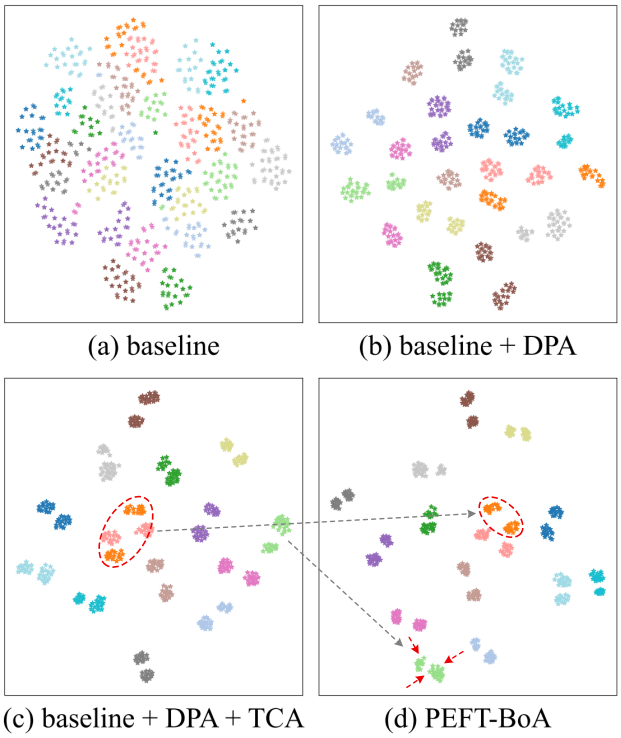


Figure 4: Feature distributions with T-SNE (Maaten and Hinton 2008). Different colors represent different IDs.

out any fine-tuning. From the results, we can observe that: 1) The adaptability of directly applying the CLIP model (#1) to the multi-modal object ReID task is very low. 2) Adding the DPA (#2)/ TCA (#3)/ IFA (#4) modules separately can significantly improve the performance of the baseline (#1). 3) Combining any two modules can further improve the model performance, as shown in #5, #6, and #7. 4) The simultaneous integration of all three components (#8) achieves the best performance, which demonstrates the complementary benefits of these modules. In conclusion, these results validate that each component contributes positively to the overall performance and that their combination enables PEFT-BoA to effectively adapt the pre-trained CLIP model for multi-modal object ReID.

**Paramater Analysis.** There are two important parameters in our model. $\lambda$ balances the contribution of Cross-Entropy loss and Triplet loss while $\alpha$ controls the weighting of cross-modal feature fusion. As shown in the Fig 3, both $\lambda$ and $\alpha$ demonstrate optimal performance at specific values. For $\lambda$, peak performance is achieved at 0.25. For $\alpha$, the best results occur at 1.0.

**Complexity Analysis.** We evaluate the complexity of our PEFT-BoA in comparison with existing CLIP-based methods. All results are obtained using the publicly released codes under consistent hardware conditions. The detailed results are summarized in Table 4. We can see that, our method requires only 6.62 million tunable parameters, which is significantly fewer than other existing methods. For computational flops, our method takes about 31.5 GFLOPs, which is comparable to or lower than most existing methods. Furthermore, our method achieves the fastest training time per epoch and the lowest training memory cost among all compared methods. It also delivers the highest inference throughput, reaching 246.2 frames per second. Moreover, our method can achieve the best performance. Overall, these results demonstrate that the proposed PEFT-BoA achieves a compelling balance between efficiency and performance.

**Feature Visualization.** As shownd in Fig. 4 (a), the baseline model shows dispersed features with limited clustering of samples within the same identity. In Fig. 4 (b), integrating the DPA module noticeably improves feature compactness by reducing intra-class distances. Adding the TCA module in Fig. 4 (c) further tightens the clustering of samples belonging to the same identity. Finally, Fig. 4 (d) demonstrates that incorporating the IFA module expands the margin between different classes, significantly improving inter-class separability. These visualizations further verify the complementary contributions of each adapter in refining feature discrimination and boosting overall ReID performance.

## Conclusion

In this paper, we propose a novel Parameter-Efficient Fine-Tuning framework with Bag-of-Adapters (PEFT-BoA) for multi-modal object ReID. To the best of our knowledge, this is the first successful adaption of CLIP model for multi-modal object ReID using purely adapter-based fine-tuning. Our framework incorporates three core adapters: the Domain-specific Patch Adapter (DPA) aligning cross-modal patches, the Task-specific Class Adapter (TCA) refining identity-discriminative tokens, and the Instance-Specific Fusion Adapter (IFA) performing selective modality fusion. Extensive experiments demonstrate that PEFT-BoA achieves state-of-the-art performance while maintaining exceptional parameter efficiency.

## Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (62302013, 62272006, 62472006) and the Natural Science Foundation of Anhui Province (2308085QF220).

## References

Cao, B.; Guo, J.; Zhu, P.; and Hu, Q. 2024. Bi-directional Adapter for Multimodal Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, 927–935.

Crawford, J.; Yin, H.; McDermott, L.; and Cummings, D. 2023. UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification. *arXiv preprint arXiv:2310.18812*.

Feng, Y.; Li, J.; Xie, C.; Tan, L.; and Ji, J. 2025. Multi-Modal Object Re-identification via Sparse Mixture-of-Experts. In *Proceedings of the International Conference on Machine Learning*, 1–9.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the International Conference on Machine Learning*, volume 97, 2790–2799.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020. Multi-Spectral Vehicle Re-Identification: A Challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11345–11353.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, 1405–1413.

Li, S.; Zheng, A.; Li, C.; Tang, J.; and Luo, B. 2025. ICPL-ReID: Identity-Conditional Prompt Learning for Multi-Spectral Object Re-Identification. *IEEE Transactions on Multimedia*, 1–12.

Lin, M.; Wang, S.; Wang, X.; Tang, J.; Fu, L.; Zuo, Z.; and Sang, N. 2025. DMPT: Decoupled Modality-aware Prompt Tuning for Multi-modal Object Re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2103–2112.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Shi, L.; Zhong, B.; Liang, Q.; Hu, X.; Mo, Z.; and Song, S. 2025. Mamba Adapter: Efficient Multi-Modal Fusion for Vision-Language Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–10.

Wan, X.; Zheng, A.; Jiang, B.; Wang, B.; Li, C.; and Tang, J. 2025. UGG-ReID: Uncertainty-Guided Graph Model for Multi-Modal Object Re-Identification. In *Advances in Neural Information Processing Systems*, 1–13.

Wang, Y.; Liu, X.; Yan, T.; Liu, Y.; Zheng, A.; Zhang, P.; and Lu, H. 2025a. MambaPro: Multi-Modal Object Re-identification with Mamba Aggregation and Synergistic Prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8150–8158.

Wang, Y.; Liu, X.; Zhang, P.; Lu, H.; Tu, Z.; and Lu, H. 2024a. Top-reid: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5758–5766.

Wang, Y.; Liu, Y.; Zheng, A.; and Zhang, P. 2025b. DeMo: Decoupled Feature-Based Mixture of Experts for Multi-Modal Object Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8141–8149.

Wang, Y.; Lv, Y.; Zhang, P.; and Lu, H. 2025c. IDEA: Inverted Text with Cooperative Deformable Aggregation for Multi-modal Object Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 29701–29710.

Wang, Z.; Huang, H.; Zheng, A.; and He, R. 2024b. Heterogeneous test-time training for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5850–5858.

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022. Interact, Embed, and EnlargE: Boosting Modality-Specific Representations for Multi-Modal Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2633–2641.

Yang, X.; Huang, J. Y.; Zhou, W.; and Chen, M. 2023. Parameter-Efficient Tuning with Special Token Adaptation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 865–872.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2025. Representation Selective Coupling via Token Sparsification for Multi-Spectral Object Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4): 3633–3648.

Zhang, P.; Wang, Y.; Liu, Y.; Tu, Z.; and Lu, H. 2024. Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17117–17126.

Zhang, S.; Luo, W.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; and Zhang, Y. 2025. Prompt-Based Modality Alignment for Effective Multi-Modal Object Re-Identification. *IEEE Transactions on Image Processing*, 34: 2450–2462.

Zheng, A.; Ma, Z.; Sun, Y.; Wang, Z.; Li, C.; and Tang, J. 2025. Flare-aware cross-modal enhancement network for

multi-spectral vehicle Re-identification. *Information Fusion*, 116: 102800.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust Multi-Modality Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3529–3537.

Zheng, A.; Zhu, X.; Ma, Z.; Li, C.; Tang, J.; and Ma, J. 2023. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100: 101901.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1116–1124.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13001–13008.