ProxyTTT: Proxy-driven Test-Time Training for Multi-modal Re-identification

Aihua Zheng[1,3], **Zhaojun Liu**[2], **Xixi Wan**[3], **Chenglong Li**[1,3*], **Jin Tang**[2], **Yan Yan** [4]

[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
[2]School of Computer Science and Technology, Anhui University
[3]School of Artificial Intelligence, Anhui University
[4]School of Computer Science, University of Illinois Chicago, Chicago, USA
ahzheng214@foxmail.com, junmain.liu@foxmail.com, xixiwan11@163.com,
lcl1314@foxmail.com, tangjin@ahu.edu.cn, yyan55@uic.edu

Abstract

Multi-modal object re-identification (ReID) aims to retrieve specific targets by leveraging complementary cues from different sensing modalities. Despite recent progress, two key challenges remain: (1) the limited ability to jointly address both modality and viewpoint discrepancies, and (2) the difficulty of effectively leveraging reliable target-domain data to improve generalization. To address these challenges, we propose **Proxy**-driven **T**est-**T**ime **T**raining (**ProxyTTT**), a unified framework that enhances both multi-modal identity representation learning and model generalization. During training, we propose a Multi-Proxy Learning (MPL) mechanism to address the representation bias across different views and modalities. MPL disentangles fine-grained modality-specific and modality-common identity proxies as semantic anchors to align identity features across diverse perspectives and sensing modalities. This alignment strategy enables the model to learn robust and discriminative global identity representations under heterogeneous modality conditions. At test time, to reliably exploit target domain data, we propose Proxy Entropy-based Selective Adaptation (PESA) for test-time training. Specifically, PESA leverages the semantic structure encoded by identity proxies to estimate prediction uncertainty via entropy, and selectively adapts the model using only high-confidence samples. This selective adaptation effectively mitigates the domain shift between training and deployment environments, improving the model's generalization in real-world scenarios. Extensive experiments on four public multi-modal ReID benchmarks (RGBNT201, RGBNT100, MSVR310, and WMVeID863) demonstrate the effectiveness of ProxyTTT.

Introduction

Object re-identification (ReID) (Sun et al. 2018; Rao et al. 2021) aims to retrieve the same identity across non-overlapping camera views. While traditional methods (Zhou et al. 2019; He et al. 2021; Chen et al. 2023; Li, Sun, and Li 2023) mainly rely on RGB images, they suffer in challenging conditions such as nighttime, low illumination, or strong backlight, where RGB quality significantly deteriorates. To address this, multi-modal ReID (Zheng et al. 2023b, 2021; Li et al. 2020; Zheng et al. 2025) introduces complementary modalities like near-infrared (NIR) and thermal infrared (TIR), which provide robust signals under poor lighting. By leveraging the strengths of different modalities, multi-modal ReID enhances visual robustness across diverse conditions.

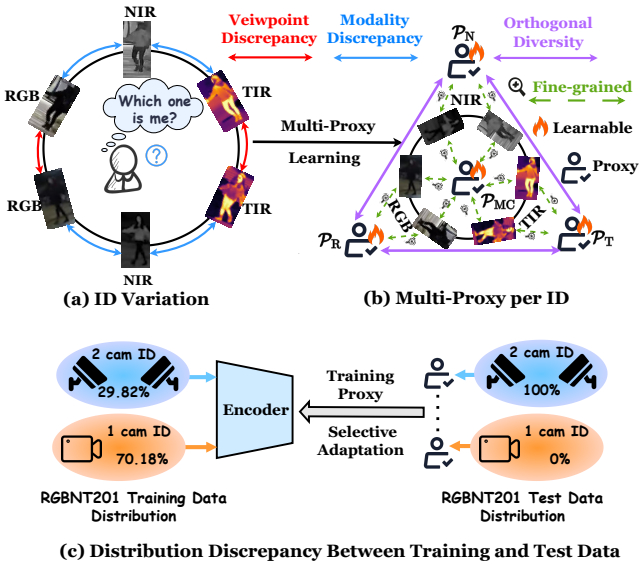

Figure 1: (a) Significant intra-identity feature variations caused by modality and viewpoint differences. (b) Our multi-proxy learning proposes both modality-specific and modality-common proxies for each identity. (c) Our proxy-based entropy estimation identifies reliable target domain samples, which are used for test-time training to reduce domain discrepancy.

In recent years, multi-modal object ReID has witnessed rapid progress. However, most existing methods (Zhang et al. 2024; Wang et al. 2025a) predominantly focus on sample-level feature interactions across modalities. For instance, Wang *et al.* (Wang et al. 2024a) proposes a token permutation strategy to explore global and local interactions between modalities at the sample level. Wang *et al* (Wang et al. 2025b) introduces a mixture-of-experts mechanism to decouple dynamic transformations of different modality fea-

*Corresponding author.

tures. Wang *et al.* (Wang et al. 2025c) is the first to incorporate an additional textual modality to enhance multi-modal feature alignment. Li *et al.* (Li et al. 2025) proposes ICPL; on the other hand, this is the first method to explore identity-level modality interactions, leveraging prototype contrastive learning to optimize text prompts for instance-level alignment. Despite these advances, current methods remain limited in scope, as they either aim to mitigate modality discrepancies at the sample level or address viewpoint discrepancies at the instance level. As illustrated in Fig. 1 (a), when an identity is simultaneously affected by both modality and viewpoint discrepancies, existing approaches struggle to resolve both challenges jointly.

Test-time training (TTT) (Sun, Darrell, and Jia 2020) has shown promise in addressing distribution shifts between the training and test sets. As illustrated in Fig. 1 (c), there exists a clear discrepancy in viewpoint distributions between the two domains. HTT (Wang et al. 2024b) is the first to introduce TTT for mitigating such distribution shifts by leveraging the classification head trained during the training phase to generate pseudo-labels for the test set, which are then used for re-training. However, we observe that relying on the classification head for pseudo-label generation introduces additional noise and fails to provide reliable supervision, ultimately limiting its effectiveness.

To address these issues, we propose ProxyTTT, a unified framework that enhances both multi-modal identity representation learning and test-time generalization for multi-modal ReID. It consists of two key components. As illustrated in Fig. 1 (b), we propose the Multi-Proxy Learning (MPL) mechanism. Unlike previous proxy-based methods (Ge et al. 2020; Huang et al. 2021; Yu et al. 2025a), our approach is the first to construct a unified multi-modal proxy learning mechanism that jointly models modality-common and modality-specific identity cues. The modality-common proxies aim to capture identity-consistent features shared across different modalities and viewpoints, thereby promoting semantic alignment and enhancing modality-invariant discrimination. In contrast, the modality-specific proxies focus on preserving complementary identity cues unique to each sensing modality, which helps retain modality diversity and enrich the feature space. To address the insufficient fine-grained attention in previous methods (Huang et al. 2021; Yu et al. 2025a), we design a fine-grained reconstruction mechanism that encourages proxies to recover local identity details, leading to more discriminative representations.

At test-time training phase (TTT), Proxy Entropy-based Selective Adaptation (PESA) enables reliable domain adaptation by leveraging the semantic structure of learned proxies. Specifically, it estimates prediction confidence based on proxy-guided entropy and selects only high-confidence samples for adaptation. This selective mechanism effectively filters noisy pseudo-labels and ensures stable optimization without requiring additional parameters. Even under partial modality degradation or missing inputs, ProxyTTT maintains strong performance. Extensive experiments on four public multi-modal ReID benchmarks demonstrate the effectiveness of ProxyTTT, consistently achieving state-of-the-art results across diverse conditions.

In summary, our contributions are as follows:

- We propose ProxyTTT, a unified framework for multi-modal ReID that jointly enhances identity learning and test-time generalization through proxy-guided representation and adaptation.
- We propose the Multi-Proxy Learning (MPL) mechanism, which simultaneously addresses both modality and viewpoint discrepancies, enabling the learning of fine-grained multi-modal identity representations.
- We design the Proxy Entropy-based Selective Adaptation (PESA) module that leverages proxy-based entropy to enable stable and effective test-time training via confident pseudo-label selection.
- Extensive experiments on four public multi-modal ReID datasets show that ProxyTTT consistently outperforms existing methods and sets new state-of-the-art results.

Related Work

Multi-modal Object ReID

In recent years, multi-modal object ReID has witnessed rapid progress. Zheng *et al*. (Li et al. 2020; Zheng et al. 2021) were among the first to introduce this task. Early works (He et al. 2023; Wang et al. 2022; Yang et al. 2025) were primarily based on convolutional neural network (CNN) (He et al. 2016; Zheng et al. 2023a; Yang et al. 2025; Wu et al. 2025), focusing on modality alignment and feature fusion. With the advent of Vision Transformer (ViT) (Alexey 2020), researchers have shown that ViT offer superior capability in modeling long-range dependencies (Crawford et al. 2023; Pan et al. 2023; Zhang et al. 2024; Yu et al. 2025b). Wang *et al.* (Wang et al. 2024a) and Wang *et al.* (Wang et al. 2024b) pioneered the integration of ViTs into multi-modal ReID, leading to notable performance improvements. More recently, with the development of CLIP (Radford et al. 2021), prompt-based approaches such as ICPL (Li et al. 2025), MambaPro (Wang et al. 2025a), and DMPT (Lin et al. 2025) have been proposed to better align heterogeneous modalities. Wang *et al.* (Wang et al. 2025b) introduces a mixture-of-experts (MoE) architecture combined with CLIP to dynamically handle modality quality variations. Wang *et al.* (Wang et al. 2025c) further enhances multi-modal representations by incorporating textual cues to enrich the semantic context of visual features.

Test-Time Training

Test-Time Training (TTT) has emerged as an effective strategy to improve model robustness under distribution shifts. Sun *et al.* (Sun, Darrell, and Jia 2020) first introduced TTT by jointly training with self-supervised and task-specific losses, using only the self-supervised objective for online adaptation. This task-agnostic framework has since been widely adopted (Han et al. 2022; Wang et al. 2021; Shin et al. 2022; Liu et al. 2021; Gandelsman et al. 2022). In multi-modal contexts, Shin *et al.* (Shin et al. 2022) proposed a TTT method based on intra-modal pseudo-labeling and inter-modal refinement, but it is limited to RGB and point

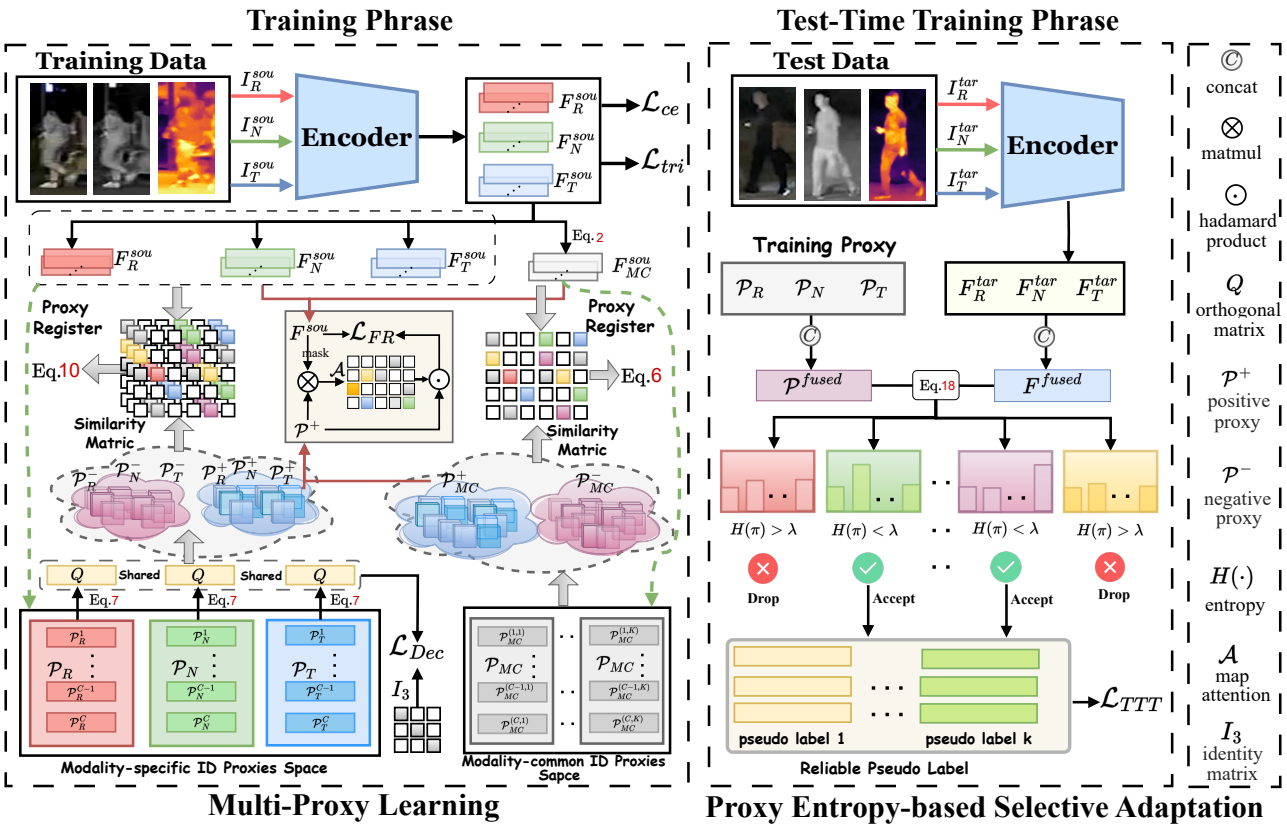

Figure 2: Overview of ProxyTTT: During training phrase, the Multi-Proxy Learning (MPL) mechanism constructs both Modality-common Proxy (MCP) and Modality-specific Proxy (MSP) to align features across modalities while preserving their unique characteristics. Fine-grained Reconstruction Learning (FRL) further enhances identity representation quality. During test-time training, the Proxy Entropy-based Selective Adaptation (PESA) module estimates uncertainty via proxy entropy and selectively adapts the model using high-confidence pseudo-labeled samples, ensuring robust generalization under domain shifts.

cloud data. Han *et al.* (Han et al. 2022) applied TTT to person ReID via BN adaptation using self-supervised losses, yet without exploiting cross-modal cues and requiring additional modules. Recently, HTT (Wang et al. 2024b) introduced the first TTT framework for multi-modal ReID using ViT backbones, but it lacks pseudo-label quality control, risking noisy adaptation. In contrast, our proposed ProxyTTT addresses this limitation through proxy-driven selective adaptation.

Methodology

As illustrated in Fig. 2, the proposed ProxyTTT framework integrates two core modules: Multi-Proxy Learning (MPL), which operates during the training phase, and Proxy Entropy-based Selective Adaptation (PESA), which is applied for test-time training phase (TTT).

Multi-Proxy Learning

To address the challenge of learning discriminative identity features under heterogeneous modality conditions, we propose MPL mechanism. It jointly models both modality-common and modality-specific identity cues by constructing corresponding proxies per identity. Fine-grained reconstruction learning enhances proxy expressiveness via local identity alignment.

Feature Extraction To effectively extract features from RGB, NIR, and TIR modalities, we adopt the shared vision encoder from pre-trained CLIP (Radford et al. 2021) as a unified feature extractor. For each modality $m \in \{R, N, T\}$, the tokenized features of source domain (training) samples are computed as follows:

$$F_m^{sou} = \Theta(I_m^{sou}) \in \mathbb{R}^D, \tag{1}$$

where $I_m^{\text{sou}} \in \mathbb{R}^{H \times W \times C}$ denotes a training image from modality m, and Θ represents the shared vision encoder. Specifically, we take the output of the `[CLS]` token from the final layer of the encoder as the global feature representation.

Modality-common Proxy (MCP). To derive a unified multi-modal representation, we concatenate the features from RGB, NIR, and TIR modalities and apply a projection for fusion and dimensionality reduction:

$$F_{\text{MC}} = \text{Proj}([F_R^{sou}, F_N^{sou}, F_T^{sou}]) \in \mathbb{R}^D, \tag{2}$$

where $[\cdot]$ denotes feature concatenation and Proj maps the fused features into a shared semantic space.

Let f_i denote the normalized modality-common feature F_{MC} of sample i. For each class c, we maintain a set of K learnable proxy vectors:

$$\mathcal{P}_c = \left\{ p_c^{(1)}, p_c^{(2)}, \ldots, p_c^{(k)}, \ldots, p_c^{(K)} \right\}, \quad p_c^{(k)} \in \mathbb{R}^D. \quad (3)$$

The positive similarity set with additive margin $\tilde{m} > 0$ is defined as the cosine similarities between the normalized feature f_i and the proxies of its ground-truth class y_i, with margin added:

$$S_i^+ = \left\{ \text{sim}(f_i, p_{y_i}^{(k)}) + \tilde{m} \;\middle|\; p_{y_i}^{(k)} \in \mathcal{P}_{y_i} \right\}. \quad (4)$$

Similarly, the negative similarity set subtracts the margin from the cosine similarities to proxies of all other classes:

$$S_i^- = \left\{ \text{sim}(f_i, p_{y_j}^{(k)}) - \tilde{m} \;\middle|\; y_j \neq y_i,\ p_{y_j}^{(k)} \in \mathcal{P}_{y_j} \right\}. \quad (5)$$

The final modality-common proxy loss is formulated as:

$$\mathcal{L}_{\text{MCP}} = \frac{1}{N} \sum_{i=1}^{N} \left[-\frac{1}{K} \sum^{S_i^+} s + \frac{1}{(C-1)K} \sum^{S_i^-} s \right], \quad (6)$$

where N is the batch size, C the number of registered classes, and K the number of proxies per class.

Modality-specific Proxy (MSP). For each modality $m \in \{R, N, T\}$, we define a set of learnable modality-specific proxies $\mathcal{P}_c^{(m)} \in \mathbb{R}^D$ for each class c. To encourage subspace disentanglement across modalities, inspired by (Wang et al. 2024c), these proxies are projected through a shared learnable orthogonal transformation matrix $Q \in \mathbb{R}^{D \times D}$:

$$\hat{\mathcal{P}}_c^{(m)} = Q\mathcal{P}_c^{(m)}, \quad \mathcal{L}_{\text{OR}} = \left\| Q^\top Q - I_D \right\|_F^2, \quad (7)$$

where $I_D \in \mathbb{R}^{D \times D}$ denotes the identity matrix.

For each class c, we form the normalized modality-specific proxy matrix by stacking the projected proxies from all modalities:

$$\hat{\mathcal{P}}_c = \left[\frac{\hat{\mathcal{P}}_c^{(R)}}{\|\hat{\mathcal{P}}_c^{(R)}\|} \quad \frac{\hat{\mathcal{P}}_c^{(N)}}{\|\hat{\mathcal{P}}_c^{(N)}\|} \quad \frac{\hat{\mathcal{P}}_c^{(T)}}{\|\hat{\mathcal{P}}_c^{(T)}\|} \right]^\top \in \mathbb{R}^{3 \times D}. \quad (8)$$

To enforce disentanglement between modality-specific subspaces, we minimize the deviation of the cross-modal similarity matrix from the identity matrix:

$$\mathcal{L}_{\text{Dec}} = \sum_{c=1}^{C} \left\| \hat{\mathcal{P}}_c (\hat{\mathcal{P}}_c)^\top - I_3 \right\|_F^2, \quad (9)$$

where $I_3 \in \mathbb{R}^{3 \times 3}$ is the 3×3 identity matrix, ensuring that the normalized proxies from different modalities for the same class are mutually orthogonal.

The modality-specific proxy loss $\mathcal{L}_{\text{SP}}^{(m)}$ aims to pull the modality-specific feature of a sample close to its class proxy, while pushing it away from proxies of other classes:

$$\mathcal{L}_{\text{SP}}^{(m)} = \frac{1}{N} \sum_{i=1}^{N} \left[-\sum^{S_i^+} s + \frac{1}{C-1} \sum^{S_i^-} s \right]. \quad (10)$$

The final modality-specific proxy loss across all modalities is aggregated as:

$$\mathcal{L}_{\text{MSP}} = \sum \mathcal{L}_{\text{SP}}^{(m)} + \mathcal{L}_{\text{Dec}} + \mathcal{L}_{\text{OR}}. \quad (11)$$

Fine-grained Reconstruction Learning (FRL). To enable the global identity proxies to capture fine-grained features, we design a Fine-grained Reconstruction Learning (FRL) mechanism. For each modality $\hat{m} \in \{R, N, T, MC\}$, we randomly mask the original feature $F_{\hat{m}}^{sou}$ to obtain the masked feature $\hat{F}_{\hat{m}}^{sou}$.

Let $P_{\hat{m}}^{pos}$ denote the set of learnable positive proxies corresponding to the identity of the sample in modality $\hat{m}$. The masked feature $\hat{F}_{\hat{m}}^{sou}$ is first passed through the simple neural network $\mathcal{X}$, and then matched against the proxy set via inner product to compute attention scores. A softmax activation is applied to produce the final attention matrix:

$$\mathcal{A} = \sigma\big(\mathcal{X}(\hat{F}_{\hat{m}}^{sou}) P_{\hat{m}}^{pos\top}\big) \in \mathbb{R}^{D \times K}, \quad (12)$$

where $\sigma(\cdot)$ denotes the softmax function, which normalizes the attention scores across the proxy dimension.

The attention weights $\mathcal{A}$ are then element-wise multiplied with $P_{\hat{m}}^{pos}$, followed by a linear layer $\mathcal{D}$, yielding the reconstructed feature:

$$\tilde{F}_{\hat{m}}^{sou} = \mathcal{D}\big(\mathcal{A} \odot P_{\hat{m}}^{pos}\big). \quad (13)$$

Finally, the reconstruction loss is computed as the mean squared error between the reconstructed feature $\tilde{F}_M$ and the original feature F_M^{sou}:

$$\mathcal{L}_{FR} = \sum_{\hat{m}} \left\| \tilde{F}_{\hat{m}}^{sou} - F_{\hat{m}}^{sou} \right\|_F^2. \quad (14)$$

This loss encourages the proxies to encode fine-grained information, improving the discriminative power of the identity representation.

Proxy Entropy-based Selective Adaptation

Test-time training (Shin et al. 2022) (TTT) has been shown to effectively mitigate the domain gap between the source and target domains. However, existing approache such as HTT (Wang et al. 2024b) struggle to fully exploit reliable target-domain data. To address this limitation, we design the Proxy Entropy-based Selective Adaptation (PESA) mechanism during TTT, which leverages source-trained modality-specific proxies to generate pseudo labels and selectively updates the model based on prediction confidence.

For each modality $m \in \{R, N, T\}$, we extract the tokenized feature from a target-domain image I_m^{tar} as:

$$F_m^{\text{tar}} = \Theta(I_m^{\text{tar}}), \quad (15)$$

where Θ denotes the shared backbone encoder. The fused multi-modal feature is obtained by concatenating modality-specific features:

$$F^{\text{fused}} = [F_R^{\text{tar}}, F_N^{\text{tar}}, F_T^{\text{tar}}] \in \mathbb{R}^{3D}. \quad (16)$$

The source-trained modality-specific proxies $\mathcal{P}_R$, $\mathcal{P}_N$, $\mathcal{P}_T \in \mathbb{R}^{C \times D}$ are concatenated into a global proxy matrix:

$$\mathcal{P}^{\text{fused}} = [\mathcal{P}_R, \mathcal{P}_N, \mathcal{P}_T] \in \mathbb{R}^{C \times 3D}. \quad (17)$$

We compute the cosine similarity between the fused feature F^{fused} and each class proxy in $\mathcal{P}^{\text{fused}}$, followed by a

temperature-scaled softmax to obtain the class probability distribution:

$$\pi = \sigma\left(\text{sim}(F^{\text{fused}}, \mathcal{P}^{\text{fused}})/T\right), \tag{18}$$

where $\pi \in \mathbb{R}^C$ denotes the predicted class probability vector, $T = 0.1$ is the temperature parameter, and $\sigma(\cdot)$ represents the softmax function. Prediction confidence is measured by the entropy of the distribution:

$$\mathcal{H} = -\sum_{c=1}^{C} \pi \log \pi. \tag{19}$$

A quantile-based threshold λ is applied to select confident samples. Let $F_{\mathcal{H}}(\cdot)$ denote the ECDF (Glivenko 1933) of entropy values within a batch. The threshold is defined as:

$$\lambda = \inf\left\{h \in \mathbb{R} \mid F_{\mathcal{H}}(h) \geq \tau\right\}, \quad \tau \in (0, 1), \tag{20}$$

where we set $\tau = 0.45$ by default. The confident set is then:

$$\mathcal{M} = \left\{\text{samples with } \mathcal{H} < \lambda\right\}. \tag{21}$$

For each confident sample, the pseudo-label is assigned as: $\hat{y} = \arg\max \pi$. The final test-time training loss is computed over the confident subset:

$$\mathcal{L}_{\text{TTT}} = \frac{1}{|\mathcal{M}|} \sum \left[\mathcal{L}_{\text{ce}}(\pi, \hat{y}) + \mathcal{L}_{\text{info}}(F^{\text{fused}}, \hat{y})\right], \tag{22}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss (Krizhevsky, Sutskever, and Hinton 2012), and $\mathcal{L}_{\text{info}}$ is an InfoNCE (Oord, Li, and Vinyals 2018) computed among confident samples sharing the same pseudo-label.

Objective Function

During training phase., the features extracted from the encoder are supervised by the cross-entropy loss (Krizhevsky, Sutskever, and Hinton 2012) and the triplet loss (Hermans, Beyer, and Leibe 2017), defined as:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}. \tag{23}$$

In addition, we incorporate auxiliary losses from the MPL. The overall training objective is formulated as:

$$\mathcal{L}_{\text{Train}} = \mathcal{L} + \mathcal{L}_{\text{MCP}} + \mathcal{L}_{\text{MSP}} + \mathcal{L}_{\text{FR}}. \tag{24}$$

The final loss function used during test-time training phase. is defined in Eq. 22.

Experiments

Dataset and Evaluation Protocols

Datasets. We evaluate our method on four representative multi-modal ReID benchmarks: RGBNT201 (Zheng et al. 2021), RGBNT100 (Li et al. 2020), MSVR310 (Zheng et al. 2023c), and WMVeID863 (Zheng et al. 2025), covering both pedestrian and vehicle scenarios under diverse conditions.

Evaluation Protocols. We adopt mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-1, 5, and 10 as standard metrics to comprehensively evaluate model performance.

Implementation Details

Our model is implemented by using the PyTorch toolbox. Experiments were conducted on RTX 3090 GPU. We use random horizontal flipping, cropping, and erasing (Zhong et al. 2020) to augment the data. During training and test-time training phase, the mini-batch size is uniformly set to 64 for all datasets, with dataset-specific sampling strategies applied accordingly. In order to optimize our model, we use the Adam optimizer, the weight attenuation is $1e^{-4}$, and the learning rate is initialized to $3.5e^{-4}$ to follow the warm-up strategy of the weight attenuation. The hyperparameter K in Equation 3 is set to 3, the margin $\tilde{m}$ in Equations 4 and 5 is set to 0.3, and the masking ratio in the FRL is set to 0.15.

Comparison with State-of-the-Art Methods

Multi-modal Person ReID. As shown in Table. 1, we compare our proposed ProxyTTT with existing methods on the RGBNT201 dataset. We observe that CNN-based methods perform worse than ViT-based TOP-ReID (Wang et al. 2024a) and CLIP-based DeMo (Wang et al. 2025b), likely due to the limited capacity of CNNs to capture long-range dependencies. Moreover, CLIP-based approaches outperform pure ViT-based ones, benefiting from large-scale pre-training that improves generalization to unseen domains. Among existing methods, IDEA (Wang et al. 2025c) achieves the best performance on RGBNT201, with an mAP of 80.2% and Rank-1 accuracy of 82.1%. This improvement is largely attributed to its introduction of an additional textual modality, which enhances visual-semantic alignment. However, leveraging extra modalities such as text also increases the difficulty of learning identity-invariant representations. In contrast, even without applying TTT, our method significantly outperforms IDEA by 2.1% in mAP and 2.6% in R-1 accuracy, while relying solely on visual modalities without introducing any auxiliary modality such as text.

Multi-modal Vehicle ReID. We further validate the effectiveness of our proposed method on the three vehicle datasets. The results on RGBNT100, MSVR310, and WMVeID863 are presented in Table. 1. On RGBNT100, IDEA (Wang et al. 2025c) achieves the best performance among existing methods. On MSVR310, ICPL (Li et al. 2025) currently obtains the highest mAP of 56.9%, which can be attributed to its use of Prototype Contrastive Learning to aggregate identity-aware semantic prompts.However, our method surpasses ICPL by 6.7%, achieving an mAP of 63.6% on MSVR310, which demonstrates its superior ability to capture fine-grained features. On the WMVeID863 dataset, which features severe glare and illumination interference, our method outperforms the current state-of-the-art MambaPro (Wang et al. 2025a), achieving gains of 2.3% in mAP and 4.4% in R-1. These results further demonstrate the strong generalization ability of our method in challenging real-world scenarios.

Evaluation under Missing Modalities. We evaluate the model's robustness under missing modalities at inference, a common real-world challenge. Despite slightly lower R-1 accuracy than DeMo (Wang et al. 2025b) when RGB and TIR modalities are missing, our model achieves higher average mAP and R-1 scores by +3.8% and +3.9%, respectively,

Methods	RGBNT201				RGBNT100		MSVR310		WMVeID863			
	mAP	R-1	R-5	R-10	mAP	R-1	mAP	R-1	mAP	R-1	R-5	R-10
HAMNet (AAAI'20)	27.7	26.3	41.5	51.7	74.5	93.3	27.1	42.3	45.6	48.5	63.1	68.8
PFNet (AAAI'21)	38.5	38.9	52.0	58.4	68.1	94.1	23.5	37.4	50.1	55.9	68.7	75.1
IEEE (AAAI'22)	49.5	48.4	59.1	65.6	64.5	83.9	21.0	41.0	45.9	48.6	64.3	67.9
CCNet (INFFUS'23)	–	–	–	–	77.2	96.3	36.4	55.2	50.3	52.7	69.6	75.1
EDITOR* (CVPR'24)	66.5	68.8	82.5	89.1	82.1	96.4	39.0	49.3	65.6	73.8	80.0	82.3
HTT* (AAAI'24)	71.1	73.4	83.1	87.3	75.7	92.6	34.5	43.2	66.2	73.2	79.9	82.3
TOP-ReID* (AAAI'24)	72.3	76.6	84.7	89.4	81.2	96.4	35.9	44.6	67.7	75.3	80.8	83.5
FACENet* (INFFUS'25)	–	–	–	–	81.5	96.9	36.2	54.1	69.8	77.0	81.0	84.2
ICPL† (TMM'25)	75.1	77.4	84.2	87.9	87.0	**98.6**	56.9	**77.7**	67.2	74.0	81.3	85.6
DMPT† (WACV'25)	78.5	81.5	90.4	93.5	81.7	94.1	36.6	52.1	–	–	–	–
DeMo† (AAAI'25)	79.0	82.3	88.8	92.0	86.2	97.6	49.2	59.8	69.6	77.2	80.8	84.2
MambaPro† (AAAI'25)	78.9	83.4	89.8	91.2	83.9	94.7	47.0	56.5	70.2	77.8	83.1	86.5
IDEA† (CVPR'25)	80.2	82.1	90.0	93.3	87.2	96.5	47.0	62.4	–	–	–	–
ProxyTTT† w/o TTT	82.3	84.7	90.6	92.7	88.4	97.9	62.1	71.7	71.5	78.8	84.0	86.7
ProxyTTT† (Ours)	**85.0**	**88.5**	**92.1**	**93.7**	**89.3**	97.7	**63.6**	72.1	**72.5**	**82.2**	**85.2**	**87.0**

Table 1: Performance comparison on RGBNT201, RGBNT100, MSVR310, and WMVeID863 datasets (in %). The symbol † denotes CLIP-based methods, * denotes ViT-based methods, and the rest are CNN-based. The best and second-best results are in **bold** and underlined, respectively.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
TOP-ReID (AAAI'24)	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	**26.0**	34.1	31.7	44.4	45.4
DeMo (AAAI'25)	63.3	65.3	72.6	75.7	56.2	54.1	45.1	46.5	26.3	24.9	40.3	38.5	50.7	50.8
ProxyTTT (Ours)	**68.5**	**68.4**	**77.1**	**79.2**	**61.4**	**61.6**	**49.6**	**51.5**	**26.5**	24.2	**43.7**	**43.2**	**54.5**	**54.7**

Table 2: Evaluation on RGBNT201 under test time modality-missing conditions. "M(X)" means missing the X image modality.

Component				RGBNT201			
MCP	MSP	FRL	PESA	mAP	R-1	R-5	R-10
×	×	×	×	75.4	75.2	82.7	86.4
✓	×	×	×	77.0	77.8	86.1	88.6
✓	✓	×	×	79.8	81.2	89.2	92.6
✓	✓	✓	×	82.3	84.7	90.6	92.7
✓	✓	✓	✓	**85.0**	**88.5**	**92.1**	**93.7**

Table 3: Comparison with different components.

on RGBNT201. This demonstrates the robustness of our approach under incomplete modality conditions.

Ablation Studies

We conduct an ablation study on the RGBNT201 dataset to evaluate the contributions of each component in ProxyTTT.

Effects of Key Modules. As shown in Table 3, the use of MCP improves the mAP from 75.4% to 77.0%, indicating its effectiveness in aligning features across modalities. Adding MSP raises the mAP to 79.8%, highlighting the benefit of modeling modality-specific identity cues. FRL further improves performance to 82.3% by enhancing local feature discrimination. The integration of PESA during test-time training further enhances generalization by selectively adapting target samples based on proxy-guided entropy.

Effects of MPL. To further verify the generality and effectiveness of the proposed Multi-Proxy Learning (MPL) mechanism, we integrate MPL into two representative methods, EDITOR (Zhang et al. 2024) and DeMo (Wang et al. 2025b). As shown in Table 4, incorporating MPL into EDITOR improves mAP and R-1 accuracy by 2.8% and 3.0%, respectively. When applied to DeMo, MPL yields gains of 3.0% in mAP and 5.0% in R-1. Notably, these improvements come with minimal additional parameters, demonstrating that MPL can consistently boost performance without incurring significant model complexity.

Effects of PESA. As shown in Table 5, we present an in-depth comparison of different pseudo-labeling strategies. Head denotes the classification-based method adopted in HTT (Wang et al. 2024b). Compared to this baseline, our proxy-based approach (Proxy) yields more stable results. Furthermore, incorporating entropy-based selective filtering (Selection) improves the quality of pseudo-labels for both the proxy and classification-based methods, leading to en-

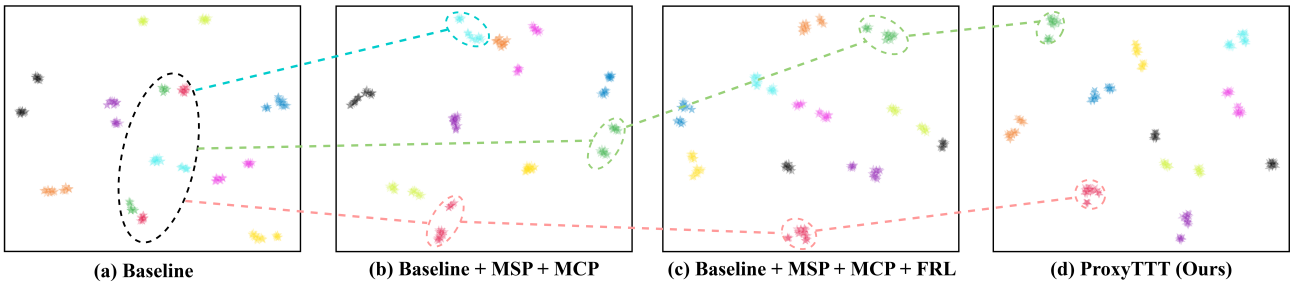

Figure 3: Feature distributions with t-SNE (Van der Maaten and Hinton 2008). Different colors refer to different IDs.

Methods	Params (M)	RGBNT201	
		mAP	R-1
EDITOR# (CVPR'24)	119.3	66.5	69.3
+ MPL	127.2	**69.3**	**72.3**
DeMo# (AAAI'25)	98.8	77.4	82.1
+ MPL	103.5	**80.4**	**87.1**

Table 4: The results of combining MPL with different methods on RGBNT201. The symbol # denotes the results replicated from previous studies.

hanced performance. These results clearly demonstrate the effectiveness of our proposed PESA strategy.

Strategy			RGBNT201			
Head	Proxy	Selection	mAP	R-1	R-5	R-10
✓	×	×	82.6	86.4	90.0	91.6
✓	×	✓	83.0	85.4	91.6	92.9
×	✓	×	83.9	86.7	92.1	93.7
×	✓	✓	**85.0**	**88.5**	**92.1**	**93.7**

Table 5: Discussion on different pseudo-labeling strategies.

Visualization Analysis

Feature Distributions. As shown in Fig. 3, our framework progressively improves feature quality. Compared to the baseline (a), incorporating MSP and MCP (b) yields better separation by modeling shared and specific features. Adding FRL (c) enhances intra-class compactness via fine-grained cues. With PESA (d), feature clusters become tighter and more distinct, highlighting the contribution of each component to discriminative representation learning.

Entropy distribution. As shown in Fig. 4, we observe that the entropy distribution of our proxy-based prediction is more left-skewed compared to the classifier-based prediction, indicating that our method produces more confident pseudo labels.

Number of Proxies. As shown in Fig. 5, we conduct experiments to evaluate the impact of varying the number of modality-common proxies K. Considering both performance and computational overhead, we select $K = 3$ as the default setting in our framework.

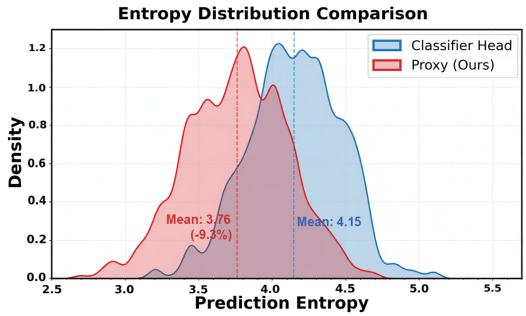

Figure 4: The entropy distribution of different pseudo-labeling methods on RGBNT201.

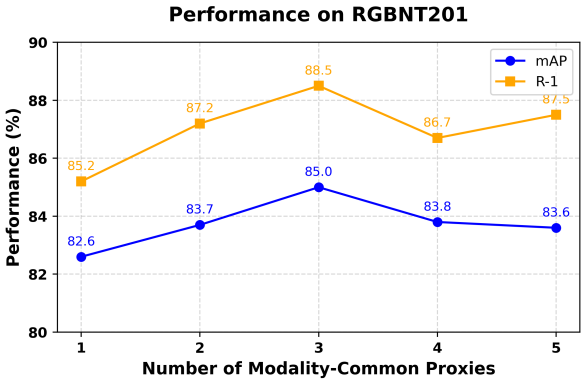

Figure 5: Number of modality-common proxies.

Conclusion

In this paper, we present a novel framework named ProxyTTT for multi-modal object ReID. During training, the MPL mechanism disentangles modality-common and modality-specific identity features via modality-common and modality-specific proxy learning, enhanced by fine-grained reconstruction. During the test-time training phase, we propose the PESA strategy, which selectively filters reliable pseudo-labeled samples to enable stable and discriminative adaptation. ProxyTTT achieves state-of-the-art performance on four challenging benchmarks.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China under Grants 62372003, the Natural Science Foundation of Anhui Province under Grants 2308085Y40

References

Alexey, D. 2020. An Image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Chen, J.; Gao, C.; Sun, L.; and Sang, N. 2023. Ccsd: Cross-camera self-distillation for unsupervised person re-identification. *Visual Intelligence*, 1(1): 27.

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

Gandelsman, Y.; Sun, Y.; Chen, X.; and Efros, A. A. 2022. Test-time Training with Masked Autoencoders. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, volume 35, 29374–29385.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in neural information processing systems*, 33: 11309–11321.

Glivenko, V. 1933. On the Empirical Determination of Laws of Probability. *Giornale dell'Istituto Italiano degli Attuari*, 4: 92–99. (in Italian).

Han, K.; Si, C.; Huang, Y.; Wang, L.; and Tan, T. 2022. Generalizable person re-identification via self-supervised batch norm test-time adaption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 817–825.

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.

He, Q.; Lu, Z.; Wang, Z.; and Hu, H. 2023. Graph-based progressive fusion network for multi-modality vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(11): 12431–12447.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 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*.

Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y.; and Zhang, Z. 2021. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11895–11904.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

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*, volume 34, 11345–11353.

Li, S.; Li, C.; Zheng, A.; Tang, J.; and Luo, B. 2025. ICPL-ReID: Identity-Conditional Prompt Learning for Multi-Spectral Object Re-Identification. *arXiv preprint arXiv:2505.17821*.

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*, volume 37, 1405–1413.

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 *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2103–2112.

Liu, Y.; Kothari, P.; Van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

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.

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 *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1025–1034.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16928–16937.

Sun, Q.; Darrell, T.; and Jia, K. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 9229–9248.

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 *Proceedings of the European conference on computer vision*, 480–496.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, D.; Zhu, X.; Yao, T.; and Darrell, T. 2021. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*.

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. 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 Computer Vision and Pattern Recognition Conference*, 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*, volume 38, 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*, volume 36, 2633–2641.

Wang, Z.; Zhao, H.; Peng, J.; Yao, L.; and Zhao, K. 2024c. Odcr: Orthogonal decoupling contrastive regularization for unpaired image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25479–25489.

Wu, D.; Liu, Z.; Chen, Z.; Gan, S.; Tan, K.; Wan, Q.; and Wang, Y. 2025. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *Expert Systems with Applications*, 263: 125716.

Yang, X.; Dong, W.; Cheng, D.; Wang, N.; and Gao, X. 2025. TIENet: A Tri-Interaction Enhancement Network for Multimodal Person Reidentification. *IEEE Transactions on Neural Networks and Learning Systems*.

Yu, C.; Liu, X.; Dai, J.; Zhang, P.; and Lu, H. 2025a. Hierarchical Proxy Learning for Cloth-Changing Person Re-Identification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2025b. 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.

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

Zheng, A.; Liu, J.; Wang, Z.; Huang, L.; Li, C.; and Yin, B. 2023b. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1(1): 29.

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*, volume 35, 3529–3537.

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

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*, volume 34, 13001–13008.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3702–3712.