JOINT MODALITY- AND VIEW-AWARE POSITIONAL RESIDUALS FOR ANY-SCENARIO PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Any-Scenario Person Re-Identification (AS-ReID) aims to match individuals across diverse combinations of sensing modalities and camera viewpoints. Although existing transformer-based methods have achieved promising performance, they commonly share positional embeddings across sensing modalities. However, this shared positional prior does not explicitly account for modality-dependent spatial cues or the changes in body-part layout caused by aerial-ground viewpoint shifts. To address these limitations, we propose Modality- and View-Aware Positional Residuals (MVPR), a framework comprising two complementary components: Modality-Aware Positional Residual Correction (MPRC) and Text-Guided View Alignment (TGVA). MPRC learns modality-specific positional residuals and injects differences between consecutive layer-specific residuals at successive transformer blocks, enabling layer-wise, position-dependent corrections within a shared visual encoder. TGVA uses learnable text prompts to supervise paired forward passes with MPRC disabled and enabled. For aerial images, the two passes use aerial- and ground-view text anchors, respectively, whereas for ground images both use ground-view anchors. These asymmetric target assignments guide aerial-to-ground adaptation while maintaining ground-view supervision for ground images. Experiments on WHU-MARS-1000 demonstrate that MVPR surpasses UAD by 3.4% and 7.7% in mAP and Rank-1 accuracy, respectively, under the AS-ReID setting. The code will be available.

1 INTRODUCTION

Person re-identification (ReID) aims to retrieve images of the same person across cameras, supporting person search in distributed camera networks (Ye et al., 2022). Any-scenario person re-identification (AS-ReID) extends this goal to arbitrary combinations of sensing modalities and camera viewpoints (Zhao et al., 2026). A single model matches individual query images against a gallery of RGB, near-infrared (NIR), and thermal infrared (TIR) observations from aerial and ground cameras. As illustrated in Fig. 1(a) and (c), spectral differences change the appearance of clothing and body regions, while aerial viewpoints substantially alter the visible body configuration. Their interaction makes identity matching across these acquisition conditions particularly challenging.

Existing transformer-based ReID methods improve representation learning through identity objectives and side information (He et al., 2021), vision-language pretraining (Radford et al., 2021), and CLIP adaptation (Li et al., 2023a). In image classification, CLIP is also adapted through prompt learning (Zhou et al., 2022; Khattak et al., 2023) and training-free classifiers (Wang et al., 2024). However, the standard Vision Transformer introduces its learned positional embeddings only at the input (Dosovitskiy et al., 2021). In the CLIP-based visual encoder, the same positional embeddings are used across modalities and added only at the encoder input, without explicit modality-specific, layer-wise positional adaptation. Fig. 1(a) illustrates this design, which provides no separate learnable positional embeddings in the subsequent layers. Spectral differences can alter the appearance and relative informativeness of identity cues across spatial regions. However, the shared positional embeddings do not explicitly account for these modality-dependent variations, and their input-only injection provides no explicit mechanism for layer-wise positional adaptation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

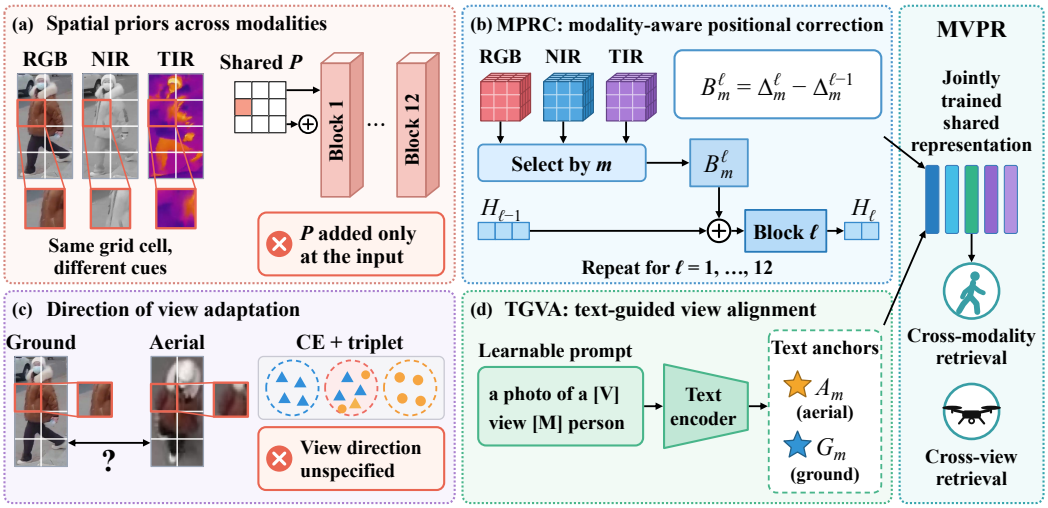


Figure 1: Illustration of the motivation of MVPR. (a) The shared positional embeddings are added only at the input, which is insufficient to handle the different visual cues at the same grid location across modalities. (b) In contrast, MPRC learns positional residuals for each modality and layer. (c) Identity objectives alone leave the direction of view alignment unspecified. (d) TGVA provides aerial and ground text anchors as explicit targets for view alignment. MVPR jointly learns a shared representation for cross-modality and cross-view retrieval within a single model.

A second difficulty concerns the direction of viewpoint adaptation. Aerial and ground cameras capture the same person with different appearances, and viewpoint changes affect identity representations (Zhang et al., 2024b; Song et al., 2025). Identity classification and triplet learning encourage cross-view identity consistency, but do not explicitly designate a reference viewpoint for positional adaptation. Modality-specific positional parameters alone likewise provide no explicit view-dependent alignment targets. As illustrated in Fig. 1(c), this absence of explicit view targets leaves the direction of positional adaptation across the aerial and ground views unspecified.

In this paper, we propose Modality- and View-Aware Positional Residuals (MVPR), a framework that combines Modality-Aware Positional Residual Correction (MPRC) with Text-Guided View Alignment (TGVA). As illustrated in Fig. 1(b), MPRC learns modality-specific positional residuals and injects the differences between consecutive layer-specific residuals before successive transformer blocks. Each image is assigned the residuals of its modality, allowing a shared visual encoder to learn position-dependent corrections for different sensing conditions. TGVA complements MPRC with view supervision from learned text prompts, as shown in Fig. 1(d). It processes each image twice through the shared visual encoder, once with MPRC disabled and once with MPRC enabled. For aerial images, the two passes are supervised by aerial- and ground-view text anchors, respectively, whereas for ground images both passes use ground-view text anchors. Together with identity objectives, these asymmetric target assignments guide aerial-to-ground adaptation in the residual-enabled pass while maintaining ground-view supervision for ground images. The main contributions are summarized as follows:

- We introduce MPRC, which learns modality-specific positional residuals throughout a shared visual encoder. By injecting the differences between adjacent residuals, it adapts the explicit positional contribution at successive transformer layers to the distinct appearance distributions of RGB, NIR, and TIR images captured by aerial and ground cameras.
- We introduce TGVA, which guides positional adaptation using learnable modality-view text prompts and paired forward passes. Its asymmetric target assignments specify aerial-to-ground alignment while retaining ground-view supervision for ground images, coupling modality-aware positional correction with an explicit reference viewpoint.
- Extensive experiments on the WHU-MARS-1000 dataset demonstrate the effectiveness of the proposed MVPR, which surpasses the previous state of the art by 3.4% and 7.7% in mAP and Rank-1 accuracy, respectively, under the AS-ReID setting.

2 Related Work

AS-ReID. Person ReID methods have made significant progress across a range of scenarios (He et al., 2024), including visible-infrared ReID (VI-ReID) scenarios (Wu et al., 2017; Nguyen et al., 2017) and aerial-ground ReID (AG-ReID) scenarios (Nguyen et al., 2024; 2025). For VI-ReID, SAAI (Fang et al., 2023) aligns latent semantic part features with learnable prototypes, DEEN (Zhang & Wang, 2023) generates diverse embeddings to reduce the modality discrepancy, and PMT (Lu et al., 2023) explores reliable identity information from modality-shared features. Recent VI-ReID methods further reduce the modality gap (Zhang et al., 2023b; Feng et al., 2023; Kim et al., 2023; Yu et al., 2023; Ren & Zhang, 2024). For AG-ReID, VDT (Zhang et al., 2024b) explicitly models and separates viewpoint-specific and viewpoint-invariant features, while GSAlign (Li et al., 2025) performs geometric transformation and visibility-aware semantic alignment. AIO (Li et al., 2024) performs all-in-one ReID across four commonly used modalities, and MixER (Alehdaghi et al., 2026) addresses mixed galleries containing both visible and infrared images. Multi-modal object ReID (Zhang et al., 2024a; Wang et al., 2025b) instead fuses RGB, NIR, and TIR images of the same object as complementary inputs. More recently, WHU-MARS extends person ReID to arbitrary combinations of modalities and viewpoints, and its UAD framework aligns modality centers toward identity centers (Zhao et al., 2026). Unlike these methods, the proposed MVPR learns modality-specific positional residuals throughout a shared visual encoder and uses textual supervision to specify the direction of cross-view adaptation within each modality.

Position embedding. In ViT, position embeddings are added to the patch embeddings to retain positional information (Dosovitskiy et al., 2021), and CLIP applies an additional layer normalization to the combined patch and position embeddings (Radford et al., 2021). CPVT (Chu et al., 2023) generates positional encodings conditioned on the local neighborhood of the input tokens. TransReID (He et al., 2021) introduces side information embeddings to mitigate feature bias towards camera and view variations. CMTR (Liang et al., 2023) fuses learnable modality embeddings with token embeddings to encode modality information. GSAlign (Li et al., 2025) performs progressive alignment through learnable thin plate spline blocks inserted before each ViT layer. Unlike embeddings added only at the input, the proposed MPRC learns positional residuals for each modality and transformer block, and TGVA supervises their cross-view adaptation with learnable text anchors.

3 Method

3.1 Overview

Fig. 2 presents the architecture of MVPR, which builds on the CLIP ViT-B/16 visual backbone (Radford et al., 2021). MVPR addresses modality-dependent feature adaptation and aerial-ground view alignment through two complementary modules. Modality-Aware Positional Residual Correction (MPRC) utilizes learnable, position-indexed corrections conditioned on the imaging modality. These corrections adapt token representations at successive transformer depths while preserving the shared backbone. Text-Guided View Alignment (TGVA) provides training supervision through two forward passes of the same image, with MPRC disabled and enabled, using six learnable text anchors corresponding to the modality-view combinations. It encourages aerial representations to align with the ground-view anchor of the same modality and retains the ground-view target for ground images. At inference, the residual-enabled visual encoder extracts query and gallery features.

Following AS-ReID (Zhao et al., 2026), each image x has an identity label y and a modality label $m \in \mathcal{M} = \{\mathrm{RGB}, \mathrm{NIR}, \mathrm{TIR}\}$. Its camera defines the viewpoint label $v \in \mathcal{V} = \{\mathrm{a}, \mathrm{g}\}$, where a and g denote aerial and ground views, respectively. The modality label m selects the positional residuals in training and inference, whereas the viewpoint label is used only by TGVA in training.

For an image divided into N non-overlapping patches, let $E(x) \in \mathbb{R}^{(N+1)\times D}$ denote the sequence of patch embeddings and a class token. We interpolate the pretrained patch positional embeddings to the input grid while retaining the class-token positional embedding. The resulting shared positional embeddings $P \in \mathbb{R}^{(N+1)\times D}$ remain trainable. Following CLIP, the initial token representation is $H_0 = \mathrm{LN}_{\mathrm{pre}}(E(x) + P)$, where $\mathrm{LN}_{\mathrm{pre}}(\cdot)$ denotes the layer normalization inherited from CLIP, which normalizes each token independently before the sequence enters the transformer blocks. The proposed MPRC injects its corrections after this normalization and before each transformer block.

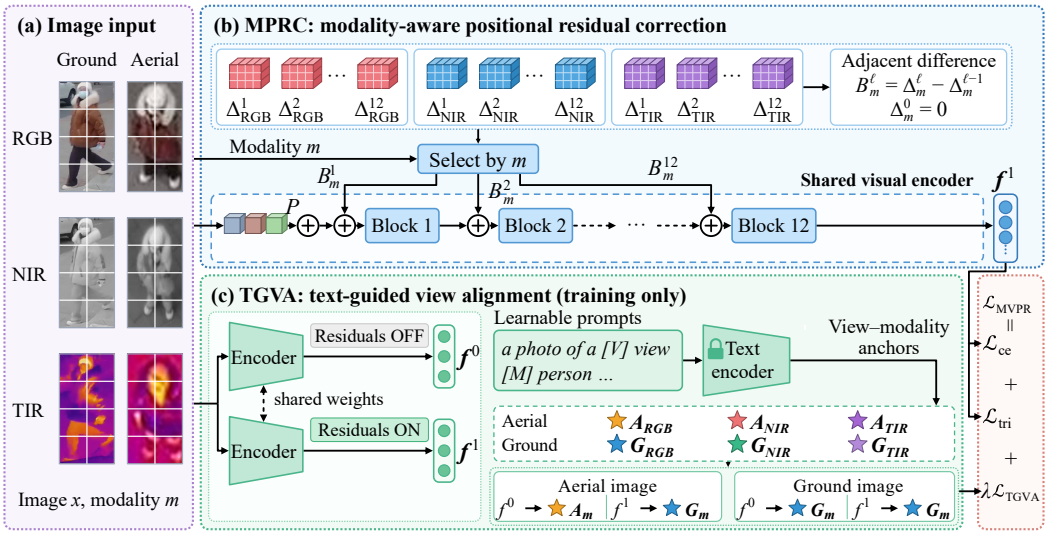


Figure 2: Overview of the proposed Modality- and View-Aware Positional Residuals (MVPR) method, which consists of a Modality-Aware Positional Residual Correction (MPRC) module and a Text-Guided View Alignment (TGVA) module. Image tokens are added to the shared positional embeddings P and layer-normalized before they enter the transformer. MPRC selects positional residuals according to the modality label m and injects the corresponding residual differences B_m^ℓ before transformer block ℓ. During training, TGVA uses residual-disabled and residual-enabled forward passes of the same image together with six learnable modality-view text anchors to supervise view alignment. Inference uses only the residual-enabled visual encoder.

The visual encoder contains L transformer blocks, indexed by $\ell = 1, \ldots, L$. Block ℓ maps an input token sequence $Z \in \mathbb{R}^{(N+1)\times D}$ to $F_\ell(Z) = Z + R_\ell(Z)$, where $R_\ell(Z)$ is the combined contribution of its multi-head self-attention and feed-forward sub-layers, each applied after layer normalization.

Applying the output layer normalization to the final class token yields the visual feature $f \in \mathbb{R}^{D\times 1}$. The BNNeck (Luo et al., 2019) output $\mathrm{BN}(f)$ is fed to a shared identity classifier. For text-guided supervision, the visual projection $W \in \mathbb{R}^{D\times d}$ maps f to the joint embedding space as $W^\top f \in \mathbb{R}^{d\times 1}$ and is fine-tuned with the visual encoder. For CLIP ViT-B/16, we have $D = 768$ and $d = 512$.

3.2 Modality-Aware Positional Residual Correction

RGB, NIR, and TIR images share the same spatial coordinate system but provide different appearance cues at each location. The shared positional embeddings P supply a common spatial prior. However, they cannot express modality-dependent adjustments to the token representations. We therefore propose MPRC to learn position-indexed feature corrections for each modality. These corrections are applied at successive transformer depths, allowing modality-specific adaptation as spatial and appearance information interact throughout feature extraction.

Specifically, MPRC learns a residual tensor $\mathbf{\Delta} \in \mathbb{R}^{|\mathcal{M}|\times L\times(N+1)\times D}$. For modality m and depth ℓ, the slice $\Delta_m^\ell \in \mathbb{R}^{(N+1)\times D}$ contains a separate correction vector for the class token and each patch position. The residuals are shared across images of the same modality. They are positional in that their entries are indexed by token location, but they act in the token feature space rather than directly replacing the input positional embeddings P. We initialize all entries of $\mathbf{\Delta}$ to zero. This makes every injected correction zero at the start of training, so enabling MPRC initially leaves the deterministic mapping of the inherited visual encoder exactly unchanged.

To define how the corrections enter the transformer, we distinguish the cumulative explicit correction at a given depth from the increment injected at that depth. Because each transformer block contains an identity path, additive injections from earlier blocks remain in the explicit additive component of later representations. Directly injecting Δ_m^ℓ at every block would therefore produce a cumulative explicit correction $\sum_{j=1}^{\ell} \Delta_m^j$ at depth ℓ. Instead, we parameterize this cumulative correction itself

by Δ_m^ℓ and inject only its change from the preceding depth:

$$B_m^\ell = \Delta_m^\ell - \Delta_m^{\ell-1}, \qquad Z_\ell = H_{\ell-1} + B_m^\ell, \qquad H_\ell = F_\ell(Z_\ell), \tag{1}$$

where $\Delta_m^0 = 0$ and $\ell = 1, \ldots, L$. Consequently, each block receives exactly the increment that changes the cumulative explicit correction at depth ℓ from $\Delta_m^{\ell-1}$ to Δ_m^ℓ.

This interpretation follows directly from the residual structure. Expanding Eq. (1) gives:

$$Z_\ell = H_0 + \Delta_m^\ell + \sum_{j=1}^{\ell-1} R_j(Z_j). \tag{2}$$

The explicit injections telescope to Δ_m^ℓ, giving each depth-specific parameter a direct interpretation as the accumulated explicit positional correction. Earlier corrections still influence the representation through the nonlinear terms $R_j(Z_j)$. Differential injection therefore controls the explicit additive component without undoing the feature transformations induced by previous corrections.

3.3 TEXT-GUIDED VIEW ALIGNMENT

MPRC provides modality-conditioned corrections, but its parameterization alone does not specify how they should address the aerial-ground view discrepancy. We introduce Text-Guided View Alignment (TGVA) to provide this guidance through modality-specific textual references. The central idea is to supervise aerial images with their original view labels when MPRC is disabled and with ground-view targets when MPRC is enabled. Ground images retain ground-view targets in both cases. This asymmetric supervision encourages the corrected features of both views to match a common ground-view reference within each modality, without viewpoint labels at inference.

Specifically, we construct a separate text anchor for each modality-view pair. Conditioning the anchors on modality allows view alignment to be supervised within RGB, NIR, and TIR, while preserving their distinct modality associations. Each prompt combines learnable view and modality embeddings with shared context vectors and fixed token sequences $E_0, \ldots, E_3$:

$$Q_{m,v} = [E_0,\, a_v,\, E_1,\, b_m,\, E_2,\, c_1, \ldots, c_K,\, E_3], \tag{3}$$

where a_v encodes viewpoint, b_m encodes modality, and the $K = 4$ context vectors $c_1, \ldots, c_K$ are shared across all six prompts. A frozen CLIP text encoder maps each prompt into a normalized text feature $t_{m,v} \in \mathbb{R}^{d\times 1}$. In Fig. 2, A_m and G_m denote $t_{m,\mathrm{a}}$ and $t_{m,\mathrm{g}}$, respectively.

For each training image x_i, we obtain features f_i^0 and f_i^1 with MPRC disabled and enabled, respectively. The two passes use the same input and share the visual encoder and projection W.

We first project both features into the joint embedding space with ℓ_2 normalization, and then compute their probability distributions over the six learnable modality-view text anchors:

$$u_i^b = \frac{W^\top f_i^b}{\left\|W^\top f_i^b\right\|_2}, \qquad p_i^b(m,v) = \frac{\exp(s\langle u_i^b, t_{m,v}\rangle)}{\sum_{m'\in\mathcal{M}}\sum_{v'\in\mathcal{V}}\exp(s\langle u_i^b, t_{m',v'}\rangle)}, \tag{4}$$

where $b \in \{0, 1\}$ indexes the two passes and s is the fixed CLIP logit scale.

For an aerial image of modality m_i, the residual-disabled and residual-enabled features are supervised with $t_{m_i,\mathrm{a}}$ and $t_{m_i,\mathrm{g}}$, respectively. The first target associates the residual-disabled feature with the observed view, while the second specifies the desired view association after correction. Together, they pull MPRC-enabled aerial features toward each modality's ground-view anchor.

For a ground image, both passes use $t_{m_i,\mathrm{g}}$, because MPRC selects corrections by modality and applies them to both aerial and ground images at inference. Supervising the residual-enabled ground features therefore helps keep the learned corrections compatible with ground-view representations, while the visual encoder and the prompt-defined anchors remain trainable.

Every target retains the image's modality label, and the softmax is taken over all six anchors, so the loss encourages both the correct modality association and the assigned view association within that modality. Let $\mathcal{B}_\mathrm{a}$ and $\mathcal{B}_\mathrm{g}$ denote the aerial and ground subsets of a training minibatch, and let v_i

Table 1: Comparison under the AS-ReID setting on WHU-MARS. Models are trained and evaluated separately on WHU-MARS-1000 and WHU-MARS-2337. In WHU-MARS-1000-GD, the WHU-MARS-1000-trained model is evaluated only on ground-view daytime queries and galleries.

Method	Venue	WHU-MARS-1000-GD				WHU-MARS-1000				WHU-MARS-2337			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
BoT (Luo et al., 2019)	CVPRW19	10.6	24.9	37.9	45.6	5.9	18.4	32.5	40.1	4.6	14.8	26.2	32.5
AGW (Ye et al., 2022)	TPAMI22	6.1	17.0	27.8	35.9	4.0	15.6	28.7	36.0	5.2	16.9	28.7	35.1
MSINet (Gu et al., 2023)	CVPR23	9.7	26.9	39.5	46.3	5.6	21.1	37.5	46.0	2.6	14.8	25.4	31.6
TransReID (He et al., 2021)	ICCV21	9.6	18.9	34.1	43.4	7.4	17.1	30.3	38.2	5.5	13.1	23.4	29.1
TransReID-SSL (Luo et al., 2021)	arXiv21	14.0	30.1	44.1	51.7	10.1	26.7	43.8	51.7	9.2	24.4	39.1	46.2
DC-Former (Li et al., 2023b)	AAAI23	9.9	20.1	34.9	42.0	6.5	16.3	29.9	36.7	4.3	10.9	19.8	25.5
CLIP-ReID (Li et al., 2023a)	AAAI23	13.8	28.9	42.7	49.6	10.6	26.6	42.6	50.4	9.3	24.1	37.5	44.0
PHA (Zhang et al., 2023a)	CVPR23	9.6	22.0	33.9	42.3	7.5	15.9	28.2	35.3	5.5	12.0	21.1	26.3
VDT (Zhang et al., 2024b)	CVPR24	13.1	31.0	44.1	51.1	7.8	25.2	41.9	50.4	6.3	20.0	33.5	40.4
DTST (Wang & Pishgar, 2025)	ICME25	11.6	29.8	42.5	47.9	6.8	23.4	39.8	47.9	7.5	23.5	37.9	45.3
SeCap (Wang et al., 2025a)	CVPR25	13.4	28.1	41.2	47.8	10.4	26.4	42.4	50.4	8.0	21.4	34.1	41.0
UAD (Zhao et al., 2026)	CVPR26	14.5	31.9	45.0	52.5	11.0	29.5	46.3	54.8	9.6	25.7	40.9	48.2
MVPR (ours)	–	**18.2**	**37.1**	**50.6**	**56.8**	**14.4**	**37.2**	**54.1**	**61.8**	**11.1**	**28.0**	**42.3**	**49.3**

denote the observed view of image i. The TGVA objective averages the two log-likelihood terms of each image over each subset:

$$\mathcal{L}_{\text{TGVA}} = -\sum_{\mathcal{S}\in\{\mathcal{B}_{\text{a}},\mathcal{B}_{\text{g}}\}} \frac{1}{\max(|\mathcal{S}|,1)} \sum_{i\in\mathcal{S}} \left[\log p_i^0(m_i, v_i) + \log p_i^1(m_i, \text{g})\right]. \tag{5}$$

3.4 Loss function and optimization

We combine TGVA with identity cross-entropy and triplet losses (Hermans et al., 2017):

$$\mathcal{L}_{\text{MVPR}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \lambda\mathcal{L}_{\text{TGVA}}, \tag{6}$$

where λ weights textual supervision. These objectives jointly optimize the positional corrections, the visual encoder, and the learnable prompts while maintaining identity-discriminative features.

4 Experiments

4.1 Experimental setup

Datasets. All experiments are conducted on the WHU-MARS dataset (Zhao et al., 2026), which contains 434,620 images of 2,337 identities captured by RGB, NIR, and TIR cameras on aerial and ground platforms across day and night. Unless otherwise specified, we use WHU-MARS-1000 as the default. Its training set contains 92,313 images of 500 identities, and its testing set contains 93,609 images of another 500 identities, from which 6,405 images are selected as queries. We additionally use the full WHU-MARS-2337 to assess scalability to more identities and scenarios.

Evaluation protocols. We report CMC (Rank-k) scores and mAP for all tasks. Following the standard VI-ReID protocol, we additionally report mean Inverse Negative Penalty (mINP) (Ye et al., 2022) in Tab. 4. We also report mINP for the ablation studies in Tab. 6 and Tab. 7.

4.2 Experimental analysis

In this section, we present a comprehensive evaluation on WHU-MARS across three settings, including AS-ReID, AG-ReID, and VI-ReID, and analyze the impact of our key design choices on performance through qualitative evaluation and systematic ablation studies.

AS-ReID. Tab. 1 compares MVPR with existing methods under the AS-ReID setting on WHU-MARS. MVPR achieves the highest mAP and Rank-1/5/10 accuracy across all three evaluation settings. On WHU-MARS-1000, it obtains 14.4% mAP and 37.2% Rank-1, outperforming UAD (Zhao et al., 2026), the strongest competing method, by 3.4% and 7.7%, respectively. The improvements extend to Rank-5 and Rank-10, indicating that MVPR places correct matches higher in the ranking list. Compared with CLIP-ReID (Li et al., 2023a), MVPR improves mAP and Rank-1 by 3.8%

and 10.6%, respectively, showing an advantage over an existing CLIP-based ReID approach under the same evaluation setting. On WHU-MARS-1000-GD, MVPR achieves 18.2% mAP and 37.1% Rank-1, exceeding UAD by 3.7% and 5.2%, respectively. On WHU-MARS-2337, where each method is trained and evaluated on that split, MVPR reaches 11.1% mAP and 28.0% Rank-1, improving over UAD by 1.5% and 2.3%, respectively. Although these margins are smaller, MVPR remains the best-performing method on the larger split across all four metrics.

Table 2: Comparison under the AG-ReID setting.

Method	A → G		G → A	
	mAP	R-1	mAP	R-1
VDT (Zhang et al., 2024b)	8.9	30.4	10.7	30.2
DTST (Wang & Pishgar, 2025)	7.9	30.5	9.8	28.5
SeCap (Wang et al., 2025a)	10.1	26.4	12.0	25.3
UAD (Zhao et al., 2026)	11.5	31.9	13.3	30.5
MVPR (ours)	**14.2**	**38.3**	**16.7**	**38.7**

Table 3: Comparison under the VI-ReID setting.

Method	NIR → RGB		TIR → RGB	
	R-1	mAP	R-1	mAP
AGW (Ye et al., 2022)	17.42	10.87	5.67	3.50
CAJ (Ye et al., 2021)	25.90	16.11	7.92	4.79
CAJ+ (Ye et al., 2024)	25.57	14.62	7.78	4.37
SAAI (Fang et al., 2023)	23.47	13.92	5.81	3.57
CSDN (Yu et al., 2025)	25.11	14.61	6.89	3.84
DEEN (Zhang & Wang, 2023)	27.63	17.05	8.01	5.69
PMT (Lu et al., 2023)	27.12	18.16	8.29	5.85
UAD (Zhao et al., 2026)	25.53	16.58	9.04	6.32
MVPR (ours)	**33.72**	**21.68**	**10.91**	**7.24**

Table 4: Modality-pair evaluation of MVPR and UAD, each using a single model. Rows and columns indicate the query and gallery modalities.

Q	G	RGB		NIR		TIR	
		MVPR	UAD	MVPR	UAD	MVPR	UAD
RGB	R-1	**48.76**	40.47	**33.91**	26.56	**9.84**	8.10
	mAP	**32.56**	26.83	**20.96**	16.52	**6.85**	6.00
	mINP	**6.78**	5.36	**2.88**	2.43	**1.12**	1.11
NIR	R-1	**33.72**	25.53	**37.05**	29.79	**10.59**	7.78
	mAP	**21.68**	16.58	**23.02**	17.93	**6.91**	5.87
	mINP	**3.65**	3.10	**3.24**	2.64	**1.19**	1.11
TIR	R-1	**10.91**	9.04	**9.51**	8.57	**26.23**	20.98
	mAP	**7.24**	6.32	**6.91**	5.69	**16.39**	14.00
	mINP	1.34	**1.36**	**1.05**	1.02	**2.81**	2.35

AG-ReID. We evaluate cross-view retrieval on WHU-MARS-1000 in two directions: aerial-to-ground (A → G) and ground-to-aerial (G → A). As shown in Tab. 2, MVPR achieves the best mAP and Rank-1 accuracy in both directions. For A → G, MVPR obtains 14.2% mAP and 38.3% Rank-1, surpassing UAD (Zhao et al., 2026) by 2.7% and 6.4%, respectively. For G → A, it reaches 16.7% mAP and 38.7% Rank-1, with corresponding gains of 3.4% and 8.2%. These improvements demonstrate the effectiveness of MVPR for retrieval across aerial and ground views in both directions. Although TGVA uses ground-view anchors as alignment targets, the resulting representation also supports retrieval with aerial images serving as the gallery set.

VI-ReID. Tab. 3 compares MVPR with existing methods on WHU-MARS-1000 under the NIR→RGB and TIR→RGB protocols. The specialized VI-ReID methods train a separate model for each modality pair, whereas MVPR uses a single model trained jointly on RGB, NIR, and TIR, without modality-pair alignment losses or VI-specific augmentations such as CAJ (Ye et al., 2021) and RLE (Tan et al., 2024). MVPR achieves the highest Rank-1 and mAP on both protocols. On NIR→RGB, it reaches 33.72% Rank-1 and 21.68% mAP, exceeding the strongest competitors in each metric, DEEN (Zhang & Wang, 2023) and PMT (Lu et al., 2023), by 6.09% and 3.52%, respectively. On TIR→RGB, MVPR obtains 10.91% Rank-1 and 7.24% mAP, improving over UAD (Zhao et al., 2026) by 1.87% and 0.92%, respectively. These results show that a single model trained for AS-ReID can outperform the compared methods on both VI-ReID tasks. Nevertheless, the low TIR→RGB performance of all methods highlights the remaining difficulty of thermal-to-visible person retrieval on the WHU-MARS-1000 dataset.

Tab. 4 further evaluates the same MVPR model across all nine query-gallery modality pairs. MVPR outperforms UAD in both Rank-1 and mAP for every pair, extending its advantage to intra-modality retrieval and the remaining cross-modality directions. Cross-modality pairs involving TIR are substantially more challenging than RGB-NIR pairs, indicating that thermal cross-modality retrieval remains a key limitation despite the consistent Rank-1 and mAP gains of MVPR over UAD. For example, MVPR achieves 33.91% Rank-1 on RGB→NIR but only 9.84% on RGB→TIR.

4.3 Ablation Studies

Progressive component ablation. We progressively add MPRC and TGVA to the CLIP baseline while keeping the initialization, training data, identity objectives, and optimization settings fixed (Tab. 5). MPRC applies modality-specific residuals to all 12 blocks, increasing mAP from 11.99% to 12.75% and Rank-1 from 31.10% to 33.47%. Adding TGVA further raises mAP to 14.42% and

Rank-1 to 37.16%, with respective gains of 1.67% and 3.69%. Rank-5 and Rank-10 also increase at both ablation steps, reaching 54.10% and 61.83% with the full model, which validates the effectiveness of both modality-specific positional correction and text-guided cross-view alignment.

Residual allocation and textual supervision. To examine how residual sharing interacts with TGVA, we compare shared and modality-specific residuals under different textual supervision (Tab. 6). Without text anchors, shared residuals achieve 13.29% mAP, higher than the 12.75% of modality-specific residuals. With six anchors, modality-specific residuals reach 14.42% and exceed shared residuals at 14.21%, with gains of 1.67% and 0.92%, respectively. Six-anchor supervision yields a larger gain for modality-specific residuals. We also examine whether to train the text encoder jointly with the prompts. Unfreezing the CLIP text encoder lowers mAP, Rank-1, Rank-5, and Rank-10 by 0.26%, 1.19%, 1.06%, and 0.91%, respectively, so we keep the text encoder frozen.

Table 5: Ablation study of the MVPR framework.

$\mathcal{L}_{\text{base}}$	MPRC	TGVA	mAP	R-1	R-5	R-10
✓			11.99	31.10	48.57	56.69
✓	✓		12.75	33.47	50.07	57.74
✓	✓	✓	**14.42**	**37.16**	**54.10**	**61.83**

Table 6: Residual allocation and textual supervision.

Residuals	Anchors	mAP	R-1	mINP
Shared	None	13.29	34.10	0.63
Shared	Two	14.11	36.14	0.64
Shared	Six	14.21	36.11	0.62
Modality-specific	None	12.75	33.47	0.64
Modality-specific	Two	14.17	35.96	**0.67**
Modality-specific	**Six**	**14.42**	**37.16**	0.65

Injection location and parameterization. We next compare injection locations with shared residuals and two RGB text anchors (Tab. 7(a)). Applying residuals to all blocks gives 14.11% mAP and 36.14% Rank-1, exceeding the best sparse results by 0.39% and 1.39%, respectively. Without text anchors, adjacent differences achieve 13.29% mAP and 34.10% Rank-1, compared with 12.76% and 32.27% for directly parameterized biases. These results show that adjacent differences outperform directly parameterized biases for the shared residuals. In the full model with modality-specific residuals and six anchors, directly parameterized biases reach 14.45% mAP and 36.89% Rank-1, compared with 14.42% and 37.16% for adjacent differences, and adjacent differences also remain slightly higher in Rank-5 and Rank-10. The advantage of adjacent differences is therefore largest without textual supervision and narrows with TGVA.

Table 7: (a) Ablation studies on the injection locations of the residuals. (b) Ablation studies on the view supervision targets. Residuals are shared and text targets use two RGB anchors on the WHU-MARS-1000 dataset. In (b), "Paired" denotes the two-pass supervision in Eq. (5).

(a) Effect of residual injection locations.

Blocks	mAP	R-1	R-5	mINP
1	13.46	34.64	51.43	0.65
12	12.83	33.51	49.95	0.58
1,4,7,10	13.72	34.66	51.98	**0.66**
1,3,5,7,9,11	13.58	34.75	51.79	0.62
All 12	**14.11**	**36.14**	**53.38**	0.64

(b) Effect of view supervision targets.

Visual parameters	Targets	mAP	R-1	R-5	mINP
Backbone	None	11.99	31.10	48.57	0.62
Backbone	Ground	12.58	32.57	49.10	0.64
Backbone + residuals	None	13.29	34.10	50.73	0.63
Backbone + residuals	Ground	13.04	34.00	50.41	**0.65**
Backbone + residuals	Paired	**14.11**	**36.14**	**53.38**	0.64

Paired view supervision. Finally, Tab. 7(b) examines whether ground-view targets alone provide effective supervision. A single forward pass with ground targets gives 12.58% mAP without residuals and 13.04% with shared residuals. Paired supervision reaches 14.11%, exceeding the single-pass shared-residual setting with two RGB text anchors by 1.07% in mAP and 2.14% in Rank-1. Ground-target accuracy reaches 100%, but retrieval remains below that of paired supervision. The additional forward pass without residuals preserves original-view targets, complementing the ground targets used with residuals. This comparison therefore supports jointly using original-view and ground-view targets in the two paired forward passes of TGVA for each training image.

4.4 Qualitative Evaluation

Retrieval examples. To further show the effectiveness of MVPR, Fig. 3 presents Rank-5 retrieval results. For each query, the retrieved images with green boxes are correct matches, while the red ones are incorrect. In general, MVPR ranks more correct matches in the top positions than the baseline, especially for the aerial RGB and NIR queries. The ground-view TIR query falls from the first to the fourth position, showing the remaining difficulty of thermal infrared matching.

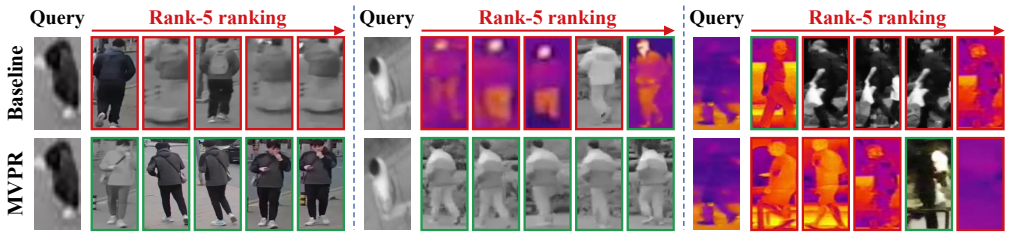


Figure 3: Rank-5 retrieval results of the baseline and the proposed MVPR on WHU-MARS-1000. Green and red boxes mark the correct and incorrect matches of each query.

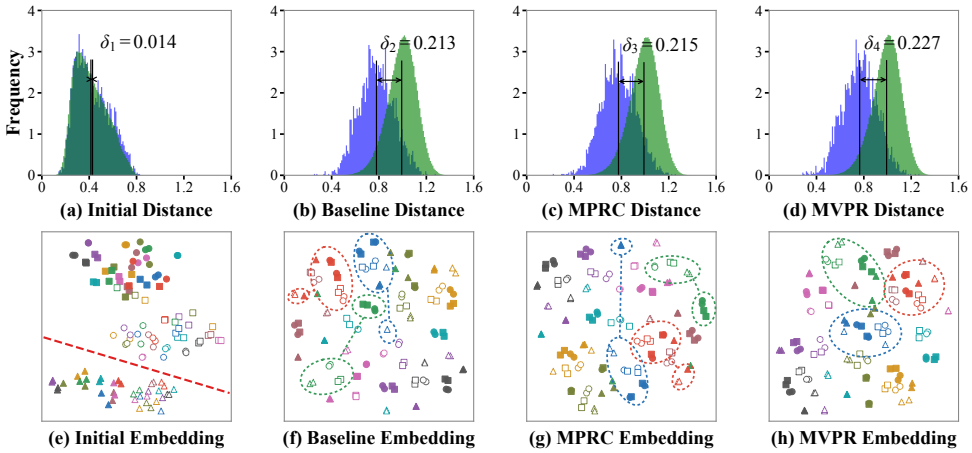


Figure 4: (a-d) show the intra-class and inter-class distances of cross-modality features. The intra-class and inter-class distances are indicated in blue and green colors, respectively. (e-h) show the distribution of feature embeddings in the 2D feature space, where circles, squares, and triangles denote the RGB, NIR, and TIR modalities, respectively. A total of 10 persons are selected from the test set. The samples with the same color are from the same person. The filled and open markers denote the images captured from the ground and aerial views, respectively.

Feature distributions. To evaluate the effectiveness of MVPR, we analyze the frequency distribution of the intra-class and inter-class cosine distances between cross-modality features in Fig. 4(a-d), following Zhang & Wang (2023). Comparing Fig. 4(b-d) with Fig. 4(a), the means (i.e., the vertical lines) of inter-class and intra-class distances are pushed away by the baseline, MPRC, and MVPR, where $\delta_1 < \delta_2 < \delta_3 < \delta_4$. Specifically, the absolute gap δ_1 is 0.014 for the original CLIP features, while δ_2, δ_3, and δ_4 are 0.213, 0.215, and 0.227, respectively. In addition, the mean intra-class distance decreases from 0.779 for the baseline features (Fig. 4(b)) to 0.765 for MVPR. Thus, MVPR can reduce the modality discrepancy among the RGB, NIR, and TIR images of the same person.

Fig. 4(e-h) shows the t-SNE embeddings (van der Maaten & Hinton, 2008). The original CLIP features are strongly organized by modality. In Fig. 4(e), the TIR samples are separated from the RGB and NIR samples by the red dashed line. Fine-tuning brings samples of the same identity closer in several groups, although some cross-modality separation remains in all three trained models.

5 Conclusion

In this paper, we propose MVPR for AS-ReID to address modality and viewpoint variations through positional adaptation. MPRC introduces modality-specific residuals at successive transformer blocks, and TGVA guides their optimization with learnable text prompts and asymmetric view targets. Experiments on WHU-MARS demonstrate improved retrieval performance across AS-ReID, AG-ReID, and VI-ReID, and progressive ablations verify the contribution of each component. On WHU-MARS-1000, MVPR surpasses UAD by 3.4% in mAP and 7.7% in Rank-1 accuracy under the AS-ReID setting. Together, positional correction and explicit view supervision enable a single model to retrieve persons across multispectral aerial and ground observations.

ETHICS STATEMENT

Person re-identification can support legitimate safety and forensic research but can also enable intrusive surveillance and identification without consent. This work uses public benchmarks, introduces no new identity data, and evaluates representation learning rather than a deployed system. Operational use would require consent and governance, human oversight, and auditing across populations, sensors, and environments. Benchmark performance alone does not establish fitness for deployment.

AI USE STATEMENT

Generative AI tools were used only to aid and polish the writing, including editing and restructuring manuscript text and formatting references and tables.The authors independently implemented the method and ran the reported experiments. They reviewed the AI-assisted material, checked the resulting claims, equations, statistics, and citations, and take full responsibility for this work.

REPRODUCIBILITY STATEMENT

Section 3 defines MPRC, TGVA, and the training objective, Section 4 specifies the datasets and evaluation protocol, and Appendix A lists the implementation details, including the input resolution, data augmentation, batch size, optimizer, learning rates, schedule, and the setting of λ. Appendix C derives the algebraic properties of the residual injection in MPRC and interprets the TGVA loss.

REFERENCES

Mahdi Alehdaghi, Rajarshi Bhattacharya, Pourya Shamsolmoali, Rafael M. O. Cruz, and Eric Granger. MixER: From Cross-Modal to Mixed-Modal Visible-Infrared Re-Identification. In *WACV*, pp. 3431–3440, 2026.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional Positional Encodings for Vision Transformers. In *ICLR*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Xingye Fang, Yang Yang, and Ying Fu. Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference. In *ICCV*, pp. 11270–11279, 2023.

Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification. In *CVPR*, pp. 22752–22761, 2023.

Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID. In *CVPR*, pp. 19243–19253, 2023.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based Object Re-Identification. In *ICCV*, pp. 15013–15022, 2021.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, Donglian Qi, and Yunfeng Yan. Instruct-ReID: A Multi-purpose Person Re-identification Task with Instructions. In *CVPR*, pp. 17521–17531, 2024.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep Networks with Stochastic Depth. In *ECCV*, pp. 646–661, 2016.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *CVPR*, pp. 19113–19122, 2023.

Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-identification. In *CVPR*, pp. 18621–18632, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

He Li, Mang Ye, Ming Zhang, and Bo Du. All in One Framework for Multimodal Re-Identification in the Wild. In *CVPR*, pp. 17459–17469, 2024.

Qiao Li, Jie Li, Yukang Zhang, Lei Tan, Jing Chen, and Jiayi Ji. GSAlign: Geometric and Semantic Alignment Network for Aerial-Ground Person Re-Identification. In *NeurIPS*, pp. 104084–104104, 2025.

Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels. In *AAAI*, pp. 1405–1413, 2023a.

Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. DC-Former: Diverse and Compact Transformer for Person Re-Identification. In *AAAI*, pp. 1415–1423, 2023b.

Tengfei Liang, Yi Jin, Wu Liu, and Yidong Li. Cross-Modality Transformer With Modality Mining for Visible-Infrared Person Re-Identification. *IEEE TMM*, 25:8432–8444, 2023.

Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning Progressive Modality-Shared Transformers for Effective Visible-Infrared Person Re-Identification. In *AAAI*, pp. 1835–1843, 2023.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPR Workshops*, pp. 1487–1495, 2019.

Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-Supervised Pre-Training for Transformer-Based Person Re-Identification. *arXiv preprint arXiv:2111.12084*, 2021.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors*, 17(3):605, 2017.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. AG-ReID.v2: Bridging Aerial and Ground Views for Person Re-identification. *IEEE TIFS*, 19:2896–2908, 2024.

Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, and Clinton Fookes. AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification. In *CVPR*, pp. 1241–1251, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pp. 8748–8763, 2021.

Kaijie Ren and Lei Zhang. Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification. In *CVPR*, pp. 393–402, 2024.

Myungseo Song, Jin-Woo Park, and Jong-Seok Lee. Exploring the Camera Bias of Person Re-identification. In *ICLR*, 2025.

Lei Tan, Yukang Zhang, Keke Han, Pingyang Dai, Yan Zhang, Yongjian Wu, and Rongrong Ji. RLE: A Unified Perspective of Data Augmentation for Cross-Spectral Re-Identification. In *NeurIPS*, pp. 126977–126996, 2024.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *JMLR*, 9:2579–2605, 2008.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. SeCap: Self-Calibrating and Adaptive Prompts for Cross-view Person Re-Identification in Aerial-Ground Networks. In *CVPR*, pp. 22119–22128, 2025a.

Yuhai Wang and Maryam Pishgar. Dynamic Token Selective Transformer for Aerial-Ground Person Re-Identification. In *ICME*, pp. 1–6, 2025.

Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. IDEA: Inverted Text with Cooperative Deformable Aggregation for Multi-modal Object Re-Identification. In *CVPR*, pp. 29701–29710, 2025b.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A Hard-to-Beat Baseline for Training-free CLIP-based Adaptation. In *ICLR*, 2024.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-Infrared Cross-Modality Person Re-Identification. In *ICCV*, pp. 5380–5389, 2017.

Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel Augmented Joint Learning for Visible-Infrared Recognition. In *ICCV*, pp. 13567–13576, 2021.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE TPAMI*, 44(6):2872–2893, 2022.

Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel Augmentation for Visible-Infrared Re-Identification. *IEEE TPAMI*, 46(4):2299–2315, 2024.

Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality Unifying Network for Visible-Infrared Person Re-Identification. In *ICCV*, pp. 11185–11195, 2023.

Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. CLIP-Driven Semantic Discovery Network for Visible-Infrared Person Re-Identification. *IEEE TMM*, 27:4137–4150, 2025.

Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. In *CVPR*, pp. 14133–14142, 2023a.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification. In *CVPR*, pp. 17117–17126, 2024a.

Quan Zhang, Lei Wang, Vishal M. Patel, Xiaohua Xie, and Jianhuang Lai. View-decoupled Transformer for Person Re-identification under Aerial-ground Camera Network. In *CVPR*, pp. 22000–22009, 2024b.

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

Yukang Zhang, Yan Yan, Jie Li, and Hanzi Wang. MRCN: A Novel Modality Restitution and Compensation Network for Visible-Infrared Person Re-identification. In *AAAI*, pp. 3498–3506, 2023b.

Yuxuan Zhao, Zhongao Zhou, Bin Yang, He Li, Jian Liang, Jun Chen, Bo Du, and Mang Ye. WHU-MARS: A Multispectral Aerial-Ground Benchmark Towards Any-Scenario Person Re-Identification. In *CVPR*, pp. 25461–25471, 2026.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. In *AAAI*, pp. 13001–13008, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *IJCV*, 130(9):2337–2348, 2022.

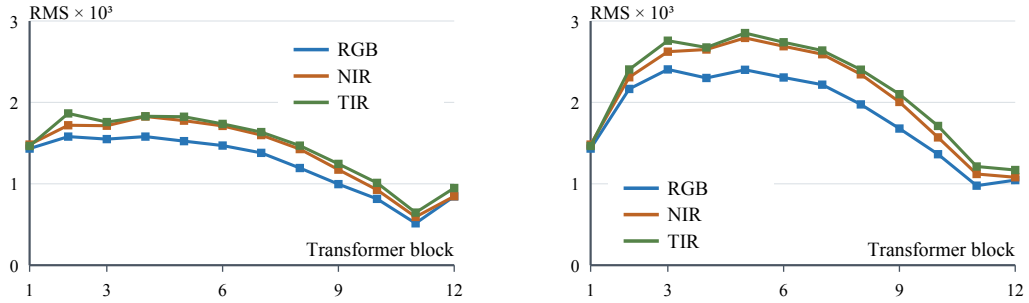
A Implementation Details

We adopt CLIP (Radford et al., 2021) as the visual and text encoders. All input images are resized to $256 \times 128 \times 3$. Random horizontal flipping, random cropping, and random erasing (Zhong et al., 2020) are adopted for data augmentation. The model is trained for 60 epochs using Adam (Kingma & Ba, 2015), and each minibatch contains 64 images per modality, with 4 images per identity. The learning rates for randomly initialized modules and pretrained components are set to 3.5×10^{-4} and 5×10^{-6}, respectively. We employ a warm-up strategy of 100 iterations, followed by a cosine learning rate schedule. Experimentally, we set $\lambda = 5$ and the weight decay to 10^{-4} by default.

The fixed subsequence E_0 contains the start-of-text token followed by "a photo of a". The subsequences E_1 and E_2 encode "view" and "person with", respectively, and E_3 contains a period and the end-of-text token. The two view embeddings are initialized from "aerial" and "ground", and the three modality embeddings are initialized from "rgb", "infrared", and "thermal". Stochastic-depth masks (Huang et al., 2016) are sampled independently for the two forward passes of TGVA.

B Learned Positional Residuals

Fig. 5 shows how MPRC allocates corrections across transformer depth. We report the root mean square (RMS), defined for a residual matrix A as $\|A\|_F/\sqrt{(N+1)D}$, averaging its squared entries over token positions and channels before taking the square root. The residual parameters Δ_m^ℓ and the injected differences B_m^ℓ vary with both modality and depth. NIR and TIR residuals are generally larger than RGB residuals, particularly in early and intermediate blocks. These results show that the learned residuals do not reduce to one correction shared across modalities and layers.



(a) Residual parameters **(b) Injected differences**

Figure 5: RMS magnitudes of (a) learned residual parameters and (b) injected adjacent differences across transformer blocks, computed from the final MVPR checkpoint on WHU-MARS-1000.

C Algebraic Properties and Loss Interpretation

Explicit additive contribution. For $\ell = 1$, Eq. (2) holds since $Z_1 = H_0 + \Delta_m^1$, and assuming that it holds at block ℓ, we obtain:

$$Z_{\ell+1} = F_\ell(Z_\ell) + \Delta_m^{\ell+1} - \Delta_m^\ell, \tag{7}$$

$$= Z_\ell + R_\ell(Z_\ell) + \Delta_m^{\ell+1} - \Delta_m^\ell, \tag{8}$$

$$= H_0 + \Delta_m^{\ell+1} + \sum_{j=1}^{\ell} R_j(Z_j). \tag{9}$$

The induction establishes cancellation only along the explicit identity connections. Since R_j depends on its input, the nonlinear transformations retain the effects of earlier residuals. The derivation holds for fixed stochastic-depth masks and does not require independence between blocks.

Equivalent additive function classes. Across all blocks, the map from Δ to B is a block lower-bidiagonal linear operator with identity diagonal and negative-identity subdiagonal, whose inverse

gives $\Delta^\ell = \sum_{j \le \ell} B^j$. Consequently, unconstrained B and Δ can implement identical visual functions when the remaining parameters agree, although the associated quadratic penalties differ:

$$\sum_\ell \|\Delta^\ell\|_F^2 = \sum_\ell \left\| \sum_{j \le \ell} B^j \right\|_F^2, \qquad \text{versus} \qquad \sum_\ell \|B^\ell\|_F^2. \tag{10}$$

Here $\|\cdot\|_F$ denotes the Frobenius norm, and this distinction, together with the transformed optimization geometry, motivates an empirical comparison rather than a proof of better optimization.

Identical passes and conflicting targets. An aerial image with identical normalized projected features in both passes contributes $-\log p_{m,\mathrm{a}} - \log p_{m,\mathrm{g}}$. Since $p_{m,\mathrm{a}} + p_{m,\mathrm{g}} \le 1$, their product is at most $1/4$, giving a lower bound of $2 \log 2$. Over the probability simplex, the minimizer assigns probability $1/2$ to each target and zero elsewhere. Finite logits reach these zero probabilities only in the limit. The two aerial terms therefore do not jointly prefer a ground target, whereas ground images retain their ground targets, and identity losses remain active. This argument assumes identical features in both passes, which independent stochastic-depth masks need not produce.

Gradients and weight decay. With $\Delta_m^\ell = \sum_{j=1}^\ell B_m^j$, the injection gradients g_B^ℓ map to $g_\Delta^\ell = g_B^\ell - g_B^{\ell+1}$ at intermediate blocks and to $g_\Delta^L = g_B^L$ at the last block, so the gradient of every intermediate residual combines those of neighboring injections. Since the depthwise stacked injections satisfy $B = S\Delta$ with the lower-bidiagonal difference operator S, gradient descent on Δ with step size η induces $B \leftarrow B - \eta S S^\top g_B$ when weight decay is omitted, whereas direct optimization gives $B \leftarrow B - \eta g_B$. Adaptive optimizers such as Adam further rescale these gradients in their respective parameter coordinates, so the two parameterizations can follow different optimization trajectories.