ProAlign: Progressive Positional and Prototype-Guided Alignment for Aerial-Ground Person Re-Identification

Qiao Li[1], **Jing Chen**[1]*, **Cong Wu**[1], **Kun He**[1] , **Ruiying Du**[1], **Lei Tan**[2]
[1] School of Cyber Science and Engineering, Wuhan University
[2]National University of Singapore
liqiaoqiao233@whu.edu.cn, chenjing@whu.edu.cn

Abstract

Aerial-Ground Person Re-Identification (AG-ReID) aims to match pedestrian images captured by unmanned aerial vehicles (UAVs) and ground-based surveillance cameras. The task remains highly challenging due to severe viewpoint discrepancies, frequent occlusions, and substantial domain gaps between aerial and ground imagery. Beyond these observable factors, existing methods often overlook two fundamental structural issues: cumulative positional drift during hierarchical feature extraction and semantic inconsistency across cross-view feature distributions. To address these challenges, we propose ProAlign, a Progressive Positional and Prototype-Guided Alignment Network for AG-ReID. ProAlign comprises two key components: a Layer-wise Progressive Positional Embedding (LPPE) module and a View-Aware Prototype Contrastive Learning (PCL) module. LPPE performs adaptive positional calibration throughout transformer layers by employing conditional positional encoding in shallow layers to mitigate local spatial distortions, while introducing learnable static positional embeddings in deeper layers to reinforce global semantic priors. Meanwhile, PCL maintains view-specific aerial and ground prototypes for each identity and enforces cross-view semantic consistency via a dual-prototype contrastive objective. Extensive experiments on CARGO and two real-world AG-ReID benchmarks demonstrate the effectiveness of ProAlign, achieving state-of-the-art overall retrieval performance on CARGO with improvements of 1.60% in Rank-1 accuracy and 5.01% in mAP over the strongest competing methods.

1 Introduction

Person re-identification (ReID) aims to match pedestrian images across distributed non-overlapping cameras and has attracted extensive attention due to its broad applications in intelligent surveillance systems. While conventional ReID has achieved remarkable progress [1, 2, 3], recent advances in unmanned aerial vehicles (UAVs) and low-altitude sensing platforms have stimulated growing interest in aerial-ground ReID (AG-ReID) [4, 5, 6]. Compared with traditional settings, AG-ReID introduces substantially larger viewpoint discrepancies and domain gaps between aerial and ground imagery, requiring models to establish robust cross-view identity correspondence under severe spatial and semantic variations.

Despite its practical importance, AG-ReID remains highly challenging and relatively underexplored. Unlike conventional ReID [7, 8, 9], aerial observations often exhibit extreme top-down viewpoints, severe occlusions, and incomplete body visibility, whereas ground images mainly contain frontal or

*Corresponding Author: Jing Chen

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

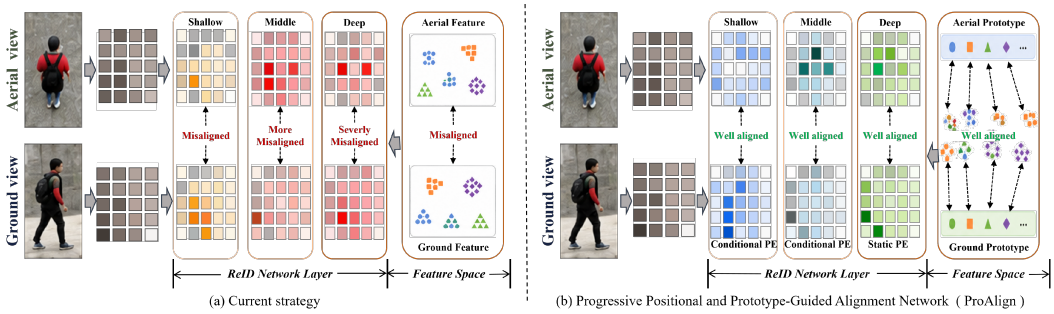

Figure 1: **Illustration of the motivation of ProAlign.** (a) Prior methods suffer from cumulative positional drift and cross-view distributional gaps. (b) ProAlign addresses these issues through layer-wise positional calibration (LPPE) and prototype-guided semantic alignment (PCL).

profile appearances. Such discrepancies introduce significant geometric deformation and appearance variation across views. More importantly, these variations are not confined to the input level. As features propagate through deep Vision Transformer (ViT) architectures, early spatial deviations progressively accumulate and evolve into hierarchical positional drift, leading to unstable high-level representations. Meanwhile, the substantial appearance gap between aerial and ground domains produces systematic semantic inconsistency across feature distributions, making robust cross-view matching particularly difficult.

Recent AG-ReID methods attempt to alleviate these issues by learning view-invariant representations or introducing geometric and semantic alignment mechanisms [10, 6, 11]. Although promising, most existing approaches primarily focus on shallow feature correction or sample-level alignment, while overlooking two fundamental structural limitations. First, positional misalignment propagates progressively across transformer hierarchies, where early spatial deviations become increasingly entangled in deeper semantic representations. Second, cross-view feature distributions remain globally inconsistent, whereas instance-level supervision alone is insufficient to preserve population-level semantic structure. Consequently, existing methods often struggle under large viewpoint discrepancies and severe visibility degradation.

To address the above limitations, we propose ProAlign, a Progressive Positional and Prototype-Guided Alignment Network for AG-ReID, as illustrated in Figure 1. The core idea of ProAlign is to jointly model hierarchical positional evolution and cross-view semantic structure within a unified framework. Specifically, we introduce a Layer-wise Progressive Positional Embedding (LPPE) module to perform depth-adaptive positional calibration throughout transformer hierarchies. LPPE employs conditional positional encoding in shallow layers to mitigate local spatial distortions, while utilizing learnable static positional embeddings in deeper layers to reinforce stable global semantic priors. This progressive design enables the network to gradually correct spatial deviations without disrupting pretrained representations. Furthermore, we propose a View-Aware Prototype Contrastive Learning (PCL) module to explicitly align aerial and ground identity distributions at the identity level. PCL maintains view-specific prototypes for each identity and updates them online through quality-aware momentum aggregation, thereby reducing the influence of noisy or occluded samples. A dual-prototype contrastive objective further encourages features from different views to converge toward shared identity-aware semantic anchors, improving cross-view consistency and discriminative capability.

To sum up, the main contributions of this paper are as follows:

- We propose ProAlign, a unified framework for aerial-ground person re-identification that jointly addresses hierarchical positional drift and cross-view semantic inconsistency under extreme viewpoint variations.
- We introduce a Layer-wise Progressive Positional Embedding (LPPE) module and a View-Aware Prototype Contrastive Learning (PCL) module. LPPE progressively calibrates spatial representations across transformer layers, while PCL performs prototype-guided population-level semantic alignment across views.

- Extensive experiments on multiple AG-ReID benchmarks demonstrate that ProAlign achieves state-of-the-art overall performance and exhibits strong robustness under challenging aerial-ground matching scenarios.

2 Related Work

Building upon well-established ground surveillance infrastructure, ReID has achieved significant progress in a variety of settings, including occluded scenarios [12, 8, 13, 14], cross-modal scenarios [15, 16, 17, 18], multi-spectral scenarios [19, 20, 21], and cross-spectral scenarios [22, 23, 24, 25, 26]. Despite these advances, most existing methods are designed under relatively stable camera viewpoints and limited cross-domain variation. With the rapid deployment of UAVs and low-altitude platforms, person ReID has been extended to aerial scenarios [27]. Early works mainly explored aerial-aerial retrieval using datasets such as UAV-Human [28] and PRAI-1581 [5]. More recently, AG-ReID has attracted increasing attention due to its practical importance. This setting introduces severe viewpoint disparities and cross-view semantic inconsistencies, leading to substantial degradation in identity-preserving features [29]. To address these challenges, recent methods explore view-invariant representation learning and cross-view alignment. Nguyen *et al.* [4, 10] propose attribute-guided multi-stream architectures. Zhang *et al.* [6] introduce a View-Decoupled Transformer to separate view-specific and view-invariant features. GSAlign [11] further models geometric deformation and visibility-aware alignment. Several recent works explicitly address geometric alignment: SeCap [30] and GeoReID [31] incorporate geometry-conditioned alignment, while GSAlign [11] models geometric deformation and visibility-aware alignment via TPS structures. Furthermore, while memory-bank and prototype-based contrastive learning have been widely explored in conventional ReID, directly adapting these instance-level and global-memory mechanisms to AG-ReID often suffers from cross-view noise and domain gaps. However, these methods primarily focus on input-level correction or instance-level alignment, while overlooking two structural issues: cumulative positional drift across transformer layers and global distributional inconsistency between aerial and ground representations. To address these limitations, we propose **ProAlign**, which jointly performs depth-adaptive positional calibration and prototype-guided cross-view alignment within a unified framework.

3 Methodology

3.1 Overview

The proposed **Progressive Positional and Prototype-Guided Alignment Network (ProAlign)**, illustrated in Fig. 2, is built upon a ViT-Base backbone for AG-ReID. It aims to address hierarchical spatial degradation and cross-view semantic inconsistency under extreme viewpoint variations. Inspired by the success of the View-Decoupled Transformer (VDT) [6], we adopt the VDT architecture as the foundation of ProAlign, which effectively captures both global semantics and fine-grained structural cues. Following its design, we introduce an additional view token to model viewpoint-specific information. The token sequence during training is defined as:

$$[\mathbf{X}_{\text{cls}}, \mathbf{X}_{\text{v}}, \mathbf{X}_{\text{img}}] = \text{VDT}\left([\mathbf{X}_{\text{cls}}, \mathbf{X}_{\text{v}}, \text{tokenize}(\mathbf{X}_{\text{img}})]\right), \tag{1}$$

where $\mathbf{X}_{\text{cls}}$, $\mathbf{X}_{\text{v}}$, and $\mathbf{X}_{\text{img}}$ represent the class token, view token, and image tokens, respectively.

To improve geometric robustness and mitigate cumulative positional drift, we integrate a **Layer-wise Progressive Positional Embedding (LPPE)** module into the transformer backbone. LPPE performs depth-adaptive positional calibration by applying conditional encodings in shallow layers and learnable static embeddings in deeper layers, progressively correcting pose-induced geometric distortions across network depth. In addition, we introduce a **View-Aware Prototype Contrastive Learning (PCL)** module during training to explicitly model identity distributions across aerial and ground views. PCL maintains quality-weighted prototypes that dynamically aggregate discriminative features, pulling samples toward consistent identity subspaces. This design effectively bridges the cross-view semantic gap while suppressing noise from background clutter and occlusions.

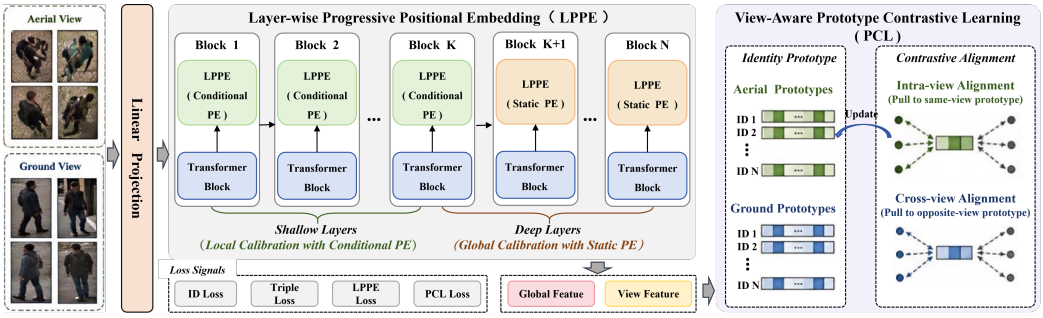

Figure 2: **Overview of the ProAlign architecture.** Given aerial or ground-view inputs, ProAlign first extracts patch tokens and processes them through transformer layers equipped with LPPE for depth-adaptive spatial correction. In parallel, the PCL module maintains view-specific prototypes to facilitate cross-view identity alignment.

3.2 Layer-wise Progressive Positional Embedding

To address spatial misalignment in person re-identification, we design a Layer-wise Progressive Positional Embedding (LPPE) module that explicitly models geometric variations induced by viewpoint discrepancies. By injecting positional corrections at each transformer layer, LPPE prevents early spatial offsets from accumulating and degrading deep representations. As illustrated in Fig. 2, LPPE is integrated into each transformer block to enable hierarchical, depth-adaptive spatial alignment. Shallow layers primarily correct local spatial distortions, while deeper layers progressively reinforce global semantic positional priors.

The core idea of LPPE is to adopt depth-dependent positional encoding strategies across transformer layers. In early stages, where spatial deviations are largely local and texture-dependent, we employ Conditional Positional Encoding (CPE), which dynamically generates positional embeddings conditioned on local patch neighborhoods via depthwise convolution. Given a patch feature map $\mathbf{F}^{(l)} \in \mathbb{R}^{B\times C\times H\times W}$, the conditional positional correction $\mathbf{E}_c^{(l)}$ is formulated as:

$$\mathbf{E}_c^{(l)} = \mathrm{DWConv}\left(\mathbf{F}^{(l)}\right), \tag{2}$$

where $\mathrm{DWConv}(\cdot)$ denotes a 3×3 depthwise convolution with zero-initialized parameters. This design enables dynamic positional correction that adapts to input-specific local spatial misalignments.

In contrast, deeper layers in ViT encode highly abstract and globally entangled representations. To stabilize these features, we introduce learnable static positional embeddings. Unlike conditional encodings, we maintain a set of independent, layer-specific embeddings $\mathbf{E}_s^{(l)} \in \mathbb{R}^{1\times N\times C}$, which serve as global semantic anchors to guide high-level representations toward a canonical spatial configuration.

To preserve pretrained representations and avoid disrupting learned semantics, we further introduce a learnable gating mechanism. The positional correction is applied only to patch tokens, while class and view tokens remain unchanged. The calibrated patch features $\tilde{\mathbf{F}}^{(l)}$ at layer l are obtained via gated residual fusion:

$$\tilde{\mathbf{F}}^{(l)} = \mathbf{F}^{(l)} + g_l \cdot \mathbf{E}^{(l)}, \tag{3}$$

where $\mathbf{E}^{(l)}$ denotes the positional embedding at layer l, selected from either $\mathbf{E}_c^{(l)}$ or $\mathbf{E}_s^{(l)}$, and g_l is a learnable scalar initialized to zero. This gating mechanism gradually activates LPPE during training, enabling a smooth introduction of spatial corrections while maintaining stability of pretrained representations under large cross-view variations.

3.3 View-Aware Prototype Contrastive Learning

In AG-ReID, severe viewpoint variations introduce a substantial distributional gap between aerial and ground representations of the same identity. Direct sample-to-sample alignment is often unreliable due to occlusions and background noise. To address this issue, we propose View-Aware Prototype Contrastive Learning (PCL), which explicitly models population-level identity distributions to bridge the cross-view semantic gap. PCL maintains a momentum-updated memory bank that anchors each

identity in a shared prototype space, thereby improving global feature consistency while reducing the influence of noisy samples.

Specifically, we construct a view-conditioned prototype memory bank during training. For each identity c, we maintain two view-specific prototypes: an aerial prototype $\mathbf{p}_c^a \in \mathbb{R}^D$ and a ground prototype $\mathbf{p}_c^g \in \mathbb{R}^D$. To enhance robustness, we introduce a quality-aware momentum update strategy that prevents occluded or low-quality samples from degrading the prototype representations. The quality weight w_i for the i-th normalized feature $\hat{\mathbf{f}}_i = \mathbf{f}_i / \|\mathbf{f}_i\|_2$ of class c in view v is defined as its cosine similarity with the corresponding prototype:

$$w_i = \max\left(0,\ \hat{\mathbf{f}}_i \cdot \mathbf{p}_c^v\right). \tag{4}$$

For each newly observed identity-view pair, the prototype is initialized using the normalized mean of its samples. For subsequent updates, nonnegative cosine similarities are used as quality weights to reduce the influence of potentially noisy features. If the sum of weights is negligible, uniform averaging is adopted to ensure a stable prototype update.

The prototypes are then updated using a momentum-based strategy. Let $\overline{\mathbf{f}}_c^v$ denote the quality-weighted mean of valid samples for class c and view v within the current mini-batch. The update rule is defined as:

$$\mathbf{p}_c^v \leftarrow \mathrm{Norm}\left((1-m)\cdot \mathbf{p}_c^v + m \cdot \overline{\mathbf{f}}_c^v\right), \tag{5}$$

where $m \in [0, 1)$ is the momentum coefficient and Norm denotes ℓ_2 normalization. For newly initialized identities, the prototype is directly set to the normalized sample mean without momentum. This weighted update scheme ensures that prototypes are dominated by discriminative and high-visibility features, improving robustness to occlusion and background noise.

During training, these prototypes serve as anchors for contrastive learning. Each sample is encouraged to align with its corresponding prototypes in both intra-view and cross-view spaces, thereby promoting viewpoint-invariant representation learning. Importantly, the prototype memory bank is used only during training to structure the feature space, and introduces no additional computational overhead during inference.

3.4 Loss Function and Optimization

Our model is trained with a composite objective that jointly optimizes identity discrimination, hierarchical spatial stability, and cross-view semantic alignment. In addition to the standard ID classification loss $\mathcal{L}_{\mathrm{id}}$ and Triplet loss $\mathcal{L}_{\mathrm{tri}}$ commonly used in ReID, we introduce a smoothness regularization loss $\mathcal{L}_{\mathrm{sm}}$ and a prototype contrastive loss $\mathcal{L}_{\mathrm{pcl}}$.

Smoothness Regularization Loss. To stabilize the static positional embeddings in deeper LPPE layers, we introduce a smoothness constraint that discourages abrupt variations between consecutive layers. This encourages gradual refinement of global positional representations. The loss is defined as the mean squared error between adjacent static embeddings:

$$\mathcal{L}_{\mathrm{sm}} = \frac{1}{|\mathcal{L}_s| - 1} \sum_{l \in \mathcal{L}_s \setminus \{\max(\mathcal{L}_s)\}} \left\| \mathbf{E}_s^{(l+1)} - \mathbf{E}_s^{(l)} \right\|_F^2. \tag{6}$$

where $\mathcal{L}_s$ denotes the set of layers equipped with static positional embeddings. This constraint promotes smooth hierarchical evolution of global spatial priors and avoids unstable layer-wise positional oscillations.

Prototype Contrastive Loss. To enforce cross-view semantic consistency at the population level, we formulate a dual-prototype InfoNCE objective $\mathcal{L}_{\mathrm{pcl}}$, consisting of intra-view and cross-view alignment terms. Given a normalized feature $\hat{\mathbf{f}}_i$ with identity label c from view $v \in \{0, 1\}$, where 0 and 1 denote ground and aerial views, respectively, the model contrasts it against all identity prototypes stored in the memory bank. The InfoNCE loss for target view t is defined as:

$$\mathcal{L}_{\mathrm{nce}}(\hat{\mathbf{f}}_i, t) = -\log \frac{\exp(\hat{\mathbf{f}}_i \cdot \mathbf{p}_c^t / \tau)}{\sum_{k=1}^{C} \exp(\hat{\mathbf{f}}_i \cdot \mathbf{p}_k^t / \tau)}, \tag{7}$$

where τ is the temperature and C is the number of identities in the memory bank.

The intra-view term enforces consistency within the same viewpoint ($t = v$), while the cross-view term aligns features across different viewpoints ($t = 1 - v$). The overall loss is formulated as:

$$\mathcal{L}_{\text{pcl}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\mathcal{L}_{\text{nce}}(\hat{\mathbf{f}}_i, v_i) + \gamma \, \mathcal{L}_{\text{nce}}(\hat{\mathbf{f}}_i, 1 - v_i) \Big). \tag{8}$$

where N denotes the batch size and γ controls the strength of cross-view alignment. This objective explicitly encourages viewpoint-invariant representation learning by aligning samples to shared identity prototypes across views.

Overall Loss. The model is trained end-to-end by minimizing a comprehensive loss function that aggregates the aforementioned components:

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}) + \alpha \, \mathcal{L}_{\text{sm}} + \beta \, \mathcal{L}_{\text{pcl}}. \tag{9}$$

where α and β balance the contributions of spatial regularization and prototype-level alignment. This joint optimization yields representations that are both spatially stable and robust to cross-view semantic discrepancies.

4 Experiments

4.1 Implementation

Datasets and Evaluation Protocols. We evaluate ProAlign on the synthetic CARGO dataset [6] and two real-world datasets, AG-ReID [4] and AG-ReID v2 [10]. CARGO provides diverse aerial and ground views under challenging viewpoint variations, resolution changes, and occlusions. We follow the official data splits and evaluation protocols for all datasets. On CARGO, we evaluate four retrieval settings: *ALL*, *G↔G*, *A↔A*, and *A↔G*. We report Rank-1, mAP, and mINP [32], except for AG-ReID v2, where mINP is not reported following prior works. All ablation and parameter studies are conducted on CARGO.

Implementation Details. All settings follow VDT [6] for fair comparison. We adopt an ImageNet-21K pretrained ViT-Base backbone with 256×128 inputs and horizontal flipping augmentation. The model is trained for 120 epochs using AdamW (weight decay 0.05) with an initial learning rate of 3.5×10^{-4}, a 20-epoch linear warm-up, and cosine decay. Each mini-batch contains $P = 64$ identities with $K = 4$ instances each. We set $\alpha = \beta = 1.0$, $\gamma = 0.5$, $m = 0.2$, and apply CPE to the first 6 layers and static embeddings to the remaining 6. All experiments are conducted on an Ubuntu 20.04 system with an Intel Xeon Gold 6133 CPU @ 2.50GHz and an NVIDIA GeForce RTX 4090 GPU. The implementation is based on Python 3.8 and PyTorch 1.12.0 with CUDA 12.0.

4.2 Comparison with State-of-the-Arts[2]

We compare ProAlign with a range of state-of-the-art person ReID methods on the CARGO dataset, including CNN-based approaches (e.g., PCB [33], BoT [34], MGN [35]), Transformer-based models (e.g., ViT [36], VDT [6], VIF [37]), and the alignment-based GSAlign [11]. As shown in Table 1, ProAlign achieves the best Rank-1, mAP and mINP under the ALL, G↔G and A↔A protocols. Under ALL, it obtains 68.27% Rank-1, 62.47% mAP and 51.39% mINP, outperforming GSAlign by +1.60%, +6.12% and +9.97%, respectively. On the challenging A↔G protocol, ProAlign improves over its VDT baseline by +12.52% Rank-1, +15.69% mAP and +19.91% mINP, and achieves the best mINP (49.86%). GSAlign obtains higher A↔G Rank-1 and mAP through explicit TPS-based geometric warping, but at the cost of 25.7M additional parameters (Table 6), whereas ProAlign adds only 1.01M and requires no geometric transformation. Notably, ProAlign also improves single-view protocols, indicating that the proposed alignment does not sacrifice intra-view retrieval.

Overall, ProAlign consistently improves retrieval performance and remains robust across diverse aerial-ground scenarios. We observe stable gains on both the CARGO dataset and real-world

[2]Extended comparisons on the AG-ReID and AG-ReID v2 datasets (Sec. A.2) are provided in the supplementary material.

Table 1: **Performance comparison of the mainstream methods under four settings of the CARGO dataset**. "ALL" denotes the overall retrieval performance of each method. "G↔G", "A↔A", and "A↔G" represent the performance of each model in several specific retrieval patterns. Rank1, mAP, and mINP are reported (%). The best performance is shown in **bold**.

Method	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [38]	50.32	43.09	29.76	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
PCB [33]	51.00	44.50	32.20	74.10	67.60	55.10	55.00	44.60	27.00	34.40	30.40	20.10
BoT [34]	54.81	46.49	32.40	77.68	66.47	51.34	65.00	49.79	29.82	36.25	32.56	21.46
MGN [35]	54.81	49.08	36.52	83.93	71.05	55.20	65.00	52.96	36.78	31.87	33.47	24.64
VV [39, 40]	45.83	38.84	39.57	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
AGW [32]	60.26	53.44	40.22	81.25	71.66	58.09	67.50	56.48	40.40	43.57	40.90	29.39
ViT [36]	61.54	53.54	39.61	82.14	71.34	57.55	80.00	64.67	47.07	43.13	40.11	28.20
GSAlign [11]	66.67	56.35	41.42	83.04	71.57	57.68	82.50	67.70	51.80	**65.96**	**60.50**	49.32
VDT [6]	64.10	55.20	41.13	82.14	71.59	58.39	82.50	66.83	50.22	48.12	42.75	29.95
VIF [37]	65.71	57.46	44.12	83.93	74.19	62.30	82.50	66.98	51.44	51.25	44.55	31.20
ProAlign	**68.27**	**62.47**	**51.39**	**84.82**	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	60.64	58.44	**49.86**

Table 2: **Ablation study of the different components in ProAlign on the CARGO dataset.** "LPPE" denotes the Layer-wise Positional Embedding module. "PCL" refers to the Prototype Contrastive Learning module. The best performance per metric is shown in **bold**.

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Baseline	64.10	55.20	41.13	82.14	71.59	58.39	82.50	66.83	50.22	48.12	42.75	29.95
Baseline + LPPE	60.58	54.92	43.13	79.46	74.10	65.12	75.00	63.62	47.28	53.19	52.63	44.73
Baseline + PCL	65.38	58.55	46.49	79.46	73.99	64.52	82.50	67.66	51.20	58.51	57.65	49.68
Baseline + LPPE + PCL	**68.27**	**62.47**	**51.39**	**84.82**	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	**60.64**	**58.44**	**49.86**

benchmarks. Results on AG-ReID and AG-ReID v2 are reported in Sec. A.2 of the supplementary material, where ProAlign also achieves consistent improvements across settings. These results indicate that the proposed positional calibration and prototype alignment mechanisms are effective and generalize across different domains.

4.3 Ablation Study

To evaluate the contribution of each component in ProAlign, we conduct an ablation study on all four protocols of the CARGO dataset. As shown in Table 2, we start from the VDT baseline [6]. **LPPE alone** yields large gains on the cross-view A↔G protocol (+5.07 Rank-1, +9.88 mAP, +14.78 mINP) and improves G↔G mAP/mINP, but it reduces Rank-1 on ALL and the intra-view protocols, as well as A↔A mAP/mINP. We attribute this to the fact that, without identity-level cross-view supervision, the newly introduced positional parameters are optimized mainly toward viewpoint-invariant spatial layouts, which slightly weakens view-specific cues useful for intra-view matching. Note that the A↔A protocol contains only 40 queries, so each query changes Rank-1 by 2.5% (the -7.50 drop corresponds to three queries). **PCL alone** improves ALL (+1.28/+3.35/+5.36) and substantially improves A↔G (+10.39/+14.90/+19.73), showing that prototype alignment is independently effective, although it slightly lowers G↔G Rank-1. **Combining both** gives the best result on every metric of every protocol, including all intra-view metrics where LPPE alone falls behind (e.g., A↔A Rank-1 from 75.00 to 85.00). This indicates that the two modules are complementary: PCL anchors the identity semantics so that LPPE's spatial calibration no longer trades off intra-view discrimination.

4.4 Discussions[3]

Depth of Conditional Positional Encoding in LPPE. We vary the number of transformer layers equipped with CPE while using static positional embeddings for the remaining layers (Table 3). We select the configuration by a predefined criterion, the best ALL-protocol mAP/mINP, rather than per-protocol tuning. Applying CPE to the first six layers achieves the best mAP/mINP on ALL, G↔G and A↔A. Using CPE only in the first three layers yields the highest ALL and G↔G

[3]A Visualization Analysis of LPPE (Sec. A.1) is provided in the supplementary material.

Table 3: **Performance comparison under different depths of CPE in LPPE.** Results are reported under four matching protocols on the CARGO dataset. The best values per column are shown in **bold**.

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Depths of CPE												
First 3 layers	**68.59**	60.97	48.7	**85.71**	77.74	67.86	80.00	68.32	53.99	60.64	60.31	52.22
First 6 layers	68.27	**62.47**	**51.39**	84.82	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	60.64	58.44	49.86
First 9 layers	66.35	60.40	48.43	81.25	76.22	67.17	82.50	68.73	53.71	**62.77**	**61.66**	**53.87**
All layers	67.63	58.87	45.41	83.93	74.92	64.11	80.00	67.58	52.16	61.70	59.60	50.12

Table 4: **Impact of the cross-view alignment weight γ in PCL.** Results are reported under four matching protocols on the CARGO dataset. The best values per column are shown in **bold**.

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
$\gamma = 0$	66.67	57.24	42.85	82.14	73.79	62.44	80.00	64.44	46.03	**63.83**	57.99	46.82
$\gamma = 0.25$	67.31	58.54	45.12	82.14	75.25	65.02	**85.00**	68.98	51.75	62.77	**60.54**	**51.29**
$\gamma = 0.5$	**68.27**	**62.47**	**51.39**	**84.82**	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	60.64	58.44	49.86
$\gamma = 0.75$	66.67	58.54	44.67	82.14	73.25	61.65	82.50	68.39	53.43	59.57	58.76	49.85

Table 5: **Comparison between different variants of prototype.** Results are reported under four matching protocols on the CARGO dataset. The best values per column are shown in **bold**.

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
Different variants of prototype												
Inner-Batch	62.82	56.12	43.55	77.68	72.57	62.64	77.50	64.44	48.94	60.64	**58.50**	49.49
Global Memory	61.86	55.40	42.71	77.68	71.83	61.10	77.50	64.53	48.89	56.38	55.80	47.80
Classifier Weights	62.82	56.30	43.98	81.25	75.06	65.08	75.00	63.90	47.79	53.19	53.52	45.54
View-Aware PCL	**68.27**	**62.47**	**51.39**	**84.82**	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	**60.64**	58.44	**49.86**

Rank-1 but lower mAP/mINP, while deeper CPE (first nine layers) gives the best A↔G results. Applying CPE to all layers yields the lowest ALL mINP (45.41%), suggesting that input-conditioned encoding throughout the network weakens overall ranking quality. These results reveal a trade-off between intra-view and cross-view retrieval across CPE depths; the first-six configuration offers the best overall balance. Rank-1 differences are partly within the granularity of CARGO's small query sets (one query changes A↔A Rank-1 by 2.5%). Notably, the full model with every CPE depth outperforms the VDT baseline on ALL mAP/mINP (Table 2), and all variants preserve spatial focus in the middle layers (Table 7).

Impact of cross-view alignment weight in PCL. The weight γ controls the trade-off between intra-view discrimination and cross-view alignment (Table 4). Without cross-view supervision ($\gamma = 0$), the model attains the highest A↔G Rank-1 but the lowest ALL mAP/mINP (57.24%/42.85%) and clearly lower A↔G mAP/mINP, indicating that it retrieves the easiest cross-view matches but ranks the remaining true matches poorly. Increasing γ to 0.5 steadily improves ALL, G↔G and A↔A, reaching 62.47% mAP and 51.39% mINP on ALL, while A↔G Rank-1 decreases moderately. Further increasing γ to 0.75 degrades ALL and G↔G performance without improving A↔G, suggesting that overly strong alignment suppresses discriminative identity cues. We therefore adopt $\gamma = 0.5$ following the same ALL-based selection criterion.

Comparison of prototype variants. We evaluate different prototype construction strategies, including Inner-Batch, Global Memory, Classifier Weights, and our View-Aware PCL. As shown in Table 5, clear performance gaps appear across all four protocols. Compared with Inner-Batch prototypes, which achieve 62.82% Rank-1 and 56.12% mAP under the ALL protocol, Global Memory slightly reduces performance to 61.86% Rank-1 and 55.40% mAP, indicating that global averaging introduces noise from occluded or low-quality samples. Classifier Weights achieve baseline performance comparable to Inner-Batch under the ALL protocol (62.82% Rank-1), but generalize poorly to cross-view retrieval, dropping to 53.19% Rank-1 on A↔G. This suggests that classification-oriented prototypes are suboptimal for cross-view discrimination. In contrast, our View-Aware PCL achieves the best results on ALL, G↔G and A↔A, improving ALL Rank-1 by +5.45% over Inner-Batch and Classifier

Table 7: **Normalized [CLS]-attention entropy $\mathcal{H}^{(l)}$ per layer** on CARGO. 1.0 denotes uniform attention over all 128 patches; lower values indicate sharper spatial focus. LPPE-r applies CPE to the first $12r$ layers (LPPE-0.5 is the default ProAlign setting).

Layer	3	4	5	6	7	8	9	10	11
VDT	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LPPE-0.25	0.982	0.982	0.977	0.951	0.881	0.851	0.864	0.857	0.986
LPPE-0.5	0.983	0.982	0.960	0.868	**0.809**	**0.839**	**0.837**	0.865	0.931
LPPE-0.75	0.982	0.985	0.958	**0.861**	0.817	0.842	0.860	0.881	0.941
LPPE-1.0	0.982	0.982	0.973	0.951	0.871	0.850	0.862	**0.845**	0.974

Weights, and G↔G mINP by +7.43% over the strongest variant. On A↔G, it matches Inner-Batch in Rank-1 and achieves the best mINP, while outperforming Global Memory and Classifier Weights by +4.26% and +7.45% Rank-1, respectively. These results indicate that single-prototype designs are insufficient for aerial-ground matching, and that decoupling view-specific prototypes mainly improves overall ranking quality and intra-view discrimination while remaining competitive in cross-view retrieval.

Computational Complexity Analysis. We analyze the computational efficiency of ProAlign in terms of trainable parameters, FLOPs, and inference latency, and compare it with VDT [6] and GSAlign [11] under the same input resolution (256×128) and hardware configuration. ProAlign introduces additional parameters of 1.01M compared to VDT, which are contributed by the Layer-wise Progressive Positional Embedding (LPPE) module, including depthwise 3×3 convolutions, layer-specific positional embeddings, and learnable gating scalars. The Prototype Contrastive Learning (PCL) module does not introduce additional trainable parameters at inference time, as the prototype memory is maintained as non-trainable buffers used only during training. In terms of computational cost, ProAlign requires 74.69 GFLOPs per forward pass. Despite the increased computation, the inference latency remains comparable, increasing by only 0.06 ms (0.54%). In contrast, GSAlign introduces 25.7M additional parameters for TPS-based warping and increases latency by 2.23%, whereas ProAlign achieves competitive alignment with substantially lower complexity and without relying on explicit geometric transformation.

Table 6: **Computational complexity comparison.** Latency is measured as the average forward pass time over 200 runs.

Setting	Params (M)	GFLOPs	Latency (ms)	ΔParams	ΔLatency
VDT	85.97	60.75	11.19	–	–
GSAlign	111.67	76.92	11.44	+29.9%	+2.23%
ProAlign	86.98	74.69	11.25	+1.17%	+0.54%

Spatial Attention Concentration. We analyze layer-wise spatial attention concentration as an indirect indicator of positional calibration. Lower normalized attention entropy indicates that the global token attends to a smaller set of spatial locations, but does not by itself establish correct spatial correspondence. We therefore complement the quantitative analysis with attention visualizations to examine whether the concentrated responses fall on identity-relevant pedestrian regions. At layer l, we average the [CLS]-to-patch attention weights over all heads and normalize them to sum to one, obtaining a distribution $\mathbf{a}^{(l)} \in \Delta^{N-1}$ over the N patch tokens ($N = 128$ for 256×128 inputs). Its normalized entropy is

$$\mathcal{H}^{(l)} = -\frac{1}{\log N} \sum_{j=1}^{N} a_j^{(l)} \log a_j^{(l)} \in [0, 1], \tag{10}$$

where $\mathcal{H}^{(l)}$=1 means perfectly uniform attention (no spatial selectivity) and lower values indicate sharper focus. Intuitively, $N^{\mathcal{H}^{(l)}}$ can be read as the effective number of attended patches: $\mathcal{H}$=1.0 corresponds to all 128 patches, while $\mathcal{H}$=0.81 corresponds to about 51 patches. Table 7 reports $\mathcal{H}^{(l)}$ averaged over paired aerial/ground test samples. The VDT baseline yields $\mathcal{H}^{(l)} \approx 1.000$ in all measured layers, i.e., its global token aggregates information almost uniformly, including background and occluded regions. In contrast, all LPPE variants reduce the entropy to 0.81–0.88 in Layers 7–10, where high-level semantics emerge, corresponding to attending to only about 40–55% of the patches regardless of CPE depth. This agrees with the attention maps in Appendix A.1 (Fig. 4). Further layer-wise analysis is provided in Appendix A.4.

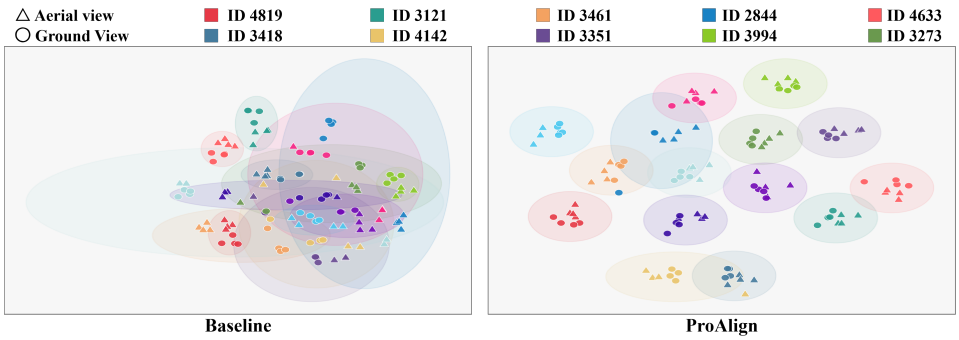

Figure 3: **t-SNE visualization of feature distributions.** △ and ◯ denote aerial and ground features, respectively. Baseline shows view-dependent clustering with fragmented identity distributions, while ProAlign produces more compact identity-wise clusters and reduces cross-view discrepancy.

Feature Distribution Analysis. We further analyze the learned feature space to evaluate the effectiveness of the proposed PCL module. Following standard practice, we extract deep features from the test set and visualize their distributions using t-SNE, as shown in Fig. 3. The baseline model exhibits a clear view-dependent clustering pattern, where samples are primarily grouped by viewpoint rather than identity. As a result, features of the same identity are scattered across different regions, indicating a pronounced cross-view semantic gap and insufficient sample-level alignment under large viewpoint variations. In contrast, ProAlign produces a more structured embedding space. Features become significantly more identity-consistent, with aerial and ground samples of the same person forming compact and unified clusters. Meanwhile, inter-identity boundaries become more distinct, suggesting improved discriminability. These observations suggest that the proposed framework improves cross-view feature consistency and identity discrimination.

5 Conclusion

In this paper, we propose ProAlign for aerial-ground person re-identification, which addresses positional drift and semantic inconsistency in cross-view scenarios. By introducing a Layer-wise Progressive Positional Embedding (LPPE) and a View-Aware Prototype Contrastive Learning (PCL), the model performs depth-adaptive positional calibration and population-level semantic alignment within a unified framework. Extensive experiments on the CARGO dataset demonstrate state-of-the-art performance across multiple protocols. We further evaluate generalization on the AG-ReID and AG-ReID v2 datasets, where consistent improvements are observed, validating the robustness of our approach under real-world cross-view variations and domain gaps.

6 Limitation and Broader Impact

Despite achieving state-of-the-art performance in AG-ReID, ProAlign still has several limitations that merit further investigation. First, the Layer-wise Progressive Positional Embedding (LPPE) relies on a predefined partition between conditional and static positional encodings across network depth. Such a fixed design lacks instance-level adaptivity and may be sub-optimal under varying degrees of spatial deformation across different samples. Second, the View-Aware Prototype Contrastive Learning (PCL) module updates prototypes using mini-batch statistics. As a result, its cross-view alignment is inherently constrained by batch composition; when an identity does not have paired samples across views within a batch, the corresponding contrastive supervision cannot be fully exploited. Future work will explore more adaptive designs, including instance-aware dynamic gating for positional calibration and cross-batch memory mechanisms to decouple prototype learning from individual batch constraints, thereby improving robustness in more complex and large-scale real-world scenarios.

Acknowledgements. This research was supported in part by the National Natural Science Foundation of China under grants No. 62625207, 62602479, 62302271, the National Cryptologic Science Fund of China under grant No. 2025NCSF02027, and the China Postdoctoral Science Foundation - Hubei Joint Support Program under Grant Number 2026T077HB.

References

[1] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[2] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1415–1423, 2023.

[3] Lei Tan, Pingyang Dai, Jie Chen, Liujuan Cao, Yongjian Wu, and Rongrong Ji. Partformer: Awakening latent diverse representation from vision transformer for object re-identification. *arXiv preprint arXiv:2408.16684*, 2024.

[4] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2585–2590. IEEE, 2023.

[5] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2020.

[6] Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22000–22009, 2024.

[7] 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. *Advances in Neural Information Processing Systems*, 37:126977–126996, 2024.

[8] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2907, 2021.

[9] Chuanming Wang, Yuxin Yang, Mengshi Qi, Huanhuan Zhang, and Huadong Ma. Towards efficient object re-identification with a novel cloud-edge collaborative framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7600–7608, 2025.

[10] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

[11] 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 *Advances in Neural Information Processing Systems*, 2025.

[12] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020.

[13] Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5070–5078, 2024.

[14] YuTeng Ye, Hang Zhou, Jiale Cai, Chenxing Gao, Youjia Zhang, Junle Wang, Qiang Hu, Junqing Yu, and Wei Yang. Dynamic feature pruning and consolidation for occluded person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6684–6692, 2024.

[15] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023.

[16] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024.

[17] Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7534–7542, 2024.

[18] Bingjun Luo, Jinpeng Wang, Zewen Wang, Junjie Zhu, and Xibin Zhao. Graph-based cross-domain knowledge distillation for cross-dataset text-to-image person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 568–576, 2025.

[19] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. 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, pages 2633–2641, 2022.

[20] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17117–17126, 2024.

[21] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8141–8149, 2025.

[22] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18621–18632, 2023.

[23] Lei Tan, Yukang Zhang, Shengmei Shen, Yan Wang, Pingyang Dai, Xianming Lin, Yongjian Wu, and Rongrong Ji. Exploring invariant representation for visible-infrared person re-identification. *arXiv preprint arXiv:2302.00884*, 2023.

[24] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Adaptive middle modality alignment learning for visible-infrared person re-identification. *International Journal of Computer Vision*, 133(4):2176–2196, 2025.

[25] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2153–2162, 2023.

[26] Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2024.

[27] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Feng Liu, Xiaoming Liu, Arun Ross, Dana Michalski, Huy Nguyen, Debayan Deb, Mahak Kothari, et al. Ag-reid 2023: Aerial-ground person re-identification challenge results. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[28] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021.

[29] Arne Schumann and Jürgen Metzler. Person re-identification across aerial and ground-based cameras by deep feature fusion. In *Automatic Target Recognition XXVII*, volume 10202, pages 56–67. SPIE, 2017.

[30] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22119–22128, 2025.

[31] Kailash A Hambarde and Hugo Proença. Rectifying geometry-induced similarity distortions for real-world aerial-ground person re-identification. *arXiv preprint arXiv:2601.21405*, 2026.

[32] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[33] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):902–917, 2019.

[34] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[35] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[37] Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9749–9758, October 2025.

[38] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9664–9667, 2023.

[39] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019.

[40] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. A strong and efficient baseline for vehicle re-identification using deep triplet embedding. *Journal of Artificial Intelligence and Soft Computing Research*, 10(1):27–45, 2020.

[41] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5056–5069, 2021.

A Supplementary Material

A.1 Visualization Analysis

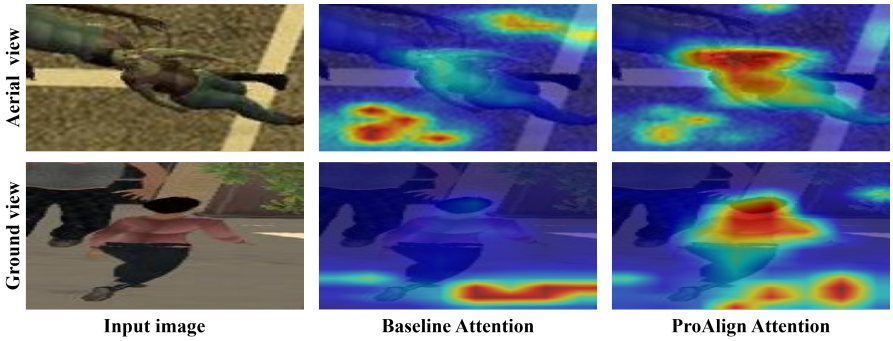

Figure 4: **Attention map comparison between baseline and ProAlign.** In both aerial and ground views, the baseline attention is scattered due to severe viewpoint distortion, while LPPE produces more focused responses on the person.

Effectiveness Analysis of the LPPE Module. We visualize attention maps to analyze the effect of the Layer-wise Progressive Positional Embedding (LPPE) module by comparing the baseline VDT with its LPPE-enhanced variant. As shown in Fig. 4, the baseline model exhibits scattered attention and frequently focuses on background regions, indicating severe spatial misalignment under large viewpoint changes. In contrast, ProAlign produces more concentrated and accurate attention distributions, consistently focusing on the target person under both aerial and ground views. This suggests that LPPE effectively improves spatial correspondence by progressively correcting positional deviations across transformer layers, leading to more stable and reliable feature localization in cross-view scenarios.

A.2 Additional Comparisons with State-of-the-Arts

Results on AG-ReID. We first evaluate ProAlign on the real-world AG-ReID dataset to assess its generalization ability. As shown in Table 8, ProAlign achieves state-of-the-art performance under the A↔G setting, reaching 86.18% Rank-1 and 77.16% mAP, outperforming strong baselines across both metrics. For the G↔A setting, VIF achieves slightly higher performance with 87.32% Rank-1 and 79.19% mAP, whereas ProAlign still maintains competitive results and achieves the best mINP with a clear margin, improving robustness under cross-view variations. Overall, these results demonstrate that ProAlign provides balanced improvements across different matching directions while preserving stable performance in real-world conditions.

Table 8: **Performance comparison of the mainstream methods under two settings of the AG-ReID dataset.** Results are reported under matching protocols. The best values per column are shown in **bold**.

Setting	Protocol 1: A↔G			Protocol 2: G↔A		
	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [38]	73.54	59.77	-	73.70	62.27	-
BoT [34]	70.01	55.47	-	71.20	58.83	-
OSNet [41]	72.59	58.32	-	74.22	60.99	-
VV [39, 40]	77.22	67.23	41.43	79.73	69.83	42.37
Explain [4]	81.47	72.61	-	82.85	73.39	-
VDT [6]	82.91	74.44	51.06	86.59	78.57	52.87
VIF [37]	83.75	75.22	51.57	**87.32**	**79.19**	52.98
ProAlign	**86.18**	**77.16**	**54.06**	85.86	78.94	**53.38**

Results on AG-ReID v2. We further evaluate ProAlign on the more challenging AG-ReID v2 dataset. As shown in Table 9, ProAlign achieves the highest Rank-1 accuracy of 91.68% under the A↔G protocol, improving upon the strongest competing result by 2.91 %. Although its Rank-1 performance is not consistently the best across all directions, ProAlign achieves the highest mAP under the G↔A, A↔W, and W↔A protocols, outperforming the strongest competing results by 7.51%, 6.65%, and 7.65 %, respectively. These results demonstrate that ProAlign effectively improves overall retrieval quality across diverse real-world matching scenarios.

A.3 Sensitivity Analysis

Momentum coefficient for prototype update. We study the sensitivity of the momentum coefficient m used in the prototype update of the PCL memory bank, which controls how rapidly prototypes adapt to new mini-batches. As shown in Table 10, we evaluate $m \in \{0.1, 0.2, 0.3\}$. Overall, $m = 0.2$ achieves the strongest overall retrieval performance, while $m = 0.1$ yields better results under the A↔G protocol. A smaller momentum

Table 9: **Performance comparison of the mainstream methods under four settings of the AG-ReID v2 dataset.** Results are reported under four matching protocols. The best values per column are shown in **bold**.

Setting	Protocol 1: A↔G		Protocol 2: G↔A		Protocol 3: A↔W		Protocol 4: W↔A	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
BoT [34]	85.40	77.03	84.65	75.90	89.77	80.48	84.65	76.90
Explain [4]	87.70	79.00	87.35	78.24	**93.67**	83.14	87.73	79.08
VDT [6]	86.46	79.13	86.14	78.12	90.00	82.21	85.26	78.52
AG-ReIDv2 [10]	88.77	80.72	87.86	78.51	93.62	84.85	**88.61**	80.11
SeCap [30]	88.12	**80.84**	**88.24**	79.99	91.44	84.01	87.56	80.15
ProAlign	**91.68**	80.17	87.19	**87.50**	91.58	**91.50**	87.09	**87.80**

Table 10: **Impact of the momentum coefficient (m) for prototype update in PCL.** Results are reported under four matching protocols on the CARGO dataset. The best values per column are shown in **bold**.

Setting	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
$m = 0.1$	67.63	60.75	47.97	81.25	75.57	66.16	80.00	67.50	51.37	**67.02**	**62.61**	**52.69**
$m = 0.2$	**68.27**	**62.47**	**51.39**	**84.82**	**80.22**	**72.51**	**85.00**	**70.05**	**54.66**	60.64	58.44	49.86
$m = 0.3$	67.31	59.24	45.95	83.93	75.79	65.32	82.50	67.00	49.45	64.89	61.42	51.48

$m = 0.1$ leads to slower prototype updates, which benefits the A↔G setting but degrades global performance in mAP and mINP. In contrast, a larger momentum $m = 0.3$ makes prototypes more sensitive to batch noise, resulting in performance degradation, particularly in the ALL setting. These results indicate that a moderate momentum effectively balances stability and adaptability, enabling robust prototype estimation under noisy and imbalanced cross-view data.

A.4 Further Analysis of Spatial Attention

This section extends the entropy analysis in Sec. 4.4. As shown in Table 7, all LPPE variants follow a consistent layer-wise pattern. In the shallow layers (3–5), the entropy remains close to uniform ($\geq$0.958) for all variants, which is expected since these layers mainly encode low-level texture rather than semantic body parts. The entropy then drops markedly in the middle layers (7–10), where semantic part representations emerge and spatial correspondence matters most, and rises again in Layer 11 (0.93–0.99), suggesting renewed attention to broader contextual information in deeper representations. The VDT baseline, by contrast, remains near-uniform in all layers, indicating that without positional calibration its global token never recovers spatial selectivity. This spatial focus is also reflected in retrieval performance: LPPE-0.5, the configuration adopted in ProAlign, attains the lowest entropy in Layers 7–9, which coincides with its best ALL mAP/mINP in Table 3, suggesting that sharper focus in the middle layers translates into better overall ranking quality.

These results also explain why LPPE calibrates only the patch tokens. The class and view tokens have no spatial coordinates and therefore cannot drift in the same sense as patch tokens; they access spatial information only through attention. Once LPPE calibrates the patch tokens, the global tokens attend to more informative regions without direct modification, as reflected by the reduced [CLS] entropy. Finally, we note that lower entropy measures how concentrated the attention is rather than whether it falls on the correct region. We therefore complement it with the attention visualizations in Fig. 4, which confirm that the concentrated regions lie on the person rather than the background.