PREDICT BEFORE YOU MATCH: CROSS-VIEW PREDICTIVE LEARNING FOR AERIAL-GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Aerial-Ground Person Re-identification (AG-ReID) task requires identity matching across observation data with significant visual feature discrepancies. Existing methods primarily address these differences through view-invariant or view-aware representation learning. This work explores a complementary approach: although the visual appearance of the same identity varies across views, predictable correlations persist between them. This implies that, beyond direct feature alignment, cross-view predictive relationships can serve as a latent supervisory signal for identity. Based on this insight, we propose an identity-conditioned cross-view latent prediction method, Predictive Re-identification across Views (PReView), designed to learn predictive relationships between observations of the same identity captured by different camera platforms. The method employs lightweight predictors to map global representations from a source view to region-level latent targets extracted from a target view. This bidirectional prediction mechanism is jointly optimized with the standard ReID objective, encouraging the encoder to learn representations that facilitate both identity discrimination and cross-view prediction. The predictors and target encoders are utilized solely during training, while the standard retrieval pipeline remains unchanged during inference. Our method demonstrates promising retrieval performance for aerial-ground person re-identification.

1 INTRODUCTION

Aerial-Ground Person Re-Identification (AG-ReID) aims to identify the same individual across heterogeneous observations captured by drones and ground-based cameras. Compared to traditional ground-based camera networks, aerial and ground platforms exhibit significant differences in shooting altitude, pitch angle, target scale, occlusion patterns, and visible body regions; consequently, the same identity presents highly inconsistent visual evidence across these distinct viewpoints. The challenge stems from more than just appearance domain shifts: the observable identity cues themselves change as viewing conditions vary. Figure 1(a) illustrates how aerial and ground observations of the same person reveal different combinations of visible identity cues. Therefore, the core issue in AG-ReID is not merely making the two viewpoints appear more similar, but rather learning representations that facilitate robust identity matching despite significant variations in visual evidence.

Existing methods primarily address this issue by focusing on the construction of more robust cross-view representations. One category of methods mitigates viewpoint bias through view disentanglement or explicit consistency constraints. For example, VDT Zhang et al. (2024) separates view-related and view-unrelated components layer by layer and utilizes the latter for identity retrieval. Some approaches posit that view-related information is not necessarily irrelevant to identity discrimination. ViSA Zhang et al. (2026) explicitly models view-invariant and view-specific semantics, fusing them through view-aware experts and local relationship modeling. GSAlign Li et al. (2026) explicitly models geometric deformation and visibility differences, improving local correspondences through geometric transformations and visibility-aware semantic masks. Another category of methods introduces additional cross-view evidence through generative completion, geometric modeling, or 3D representation. SD-ReID Wang et al. (2026) employs Stable Diffusion to generate view-specific and missing cross-view identity features, thereby supplementing identity information ab-

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.

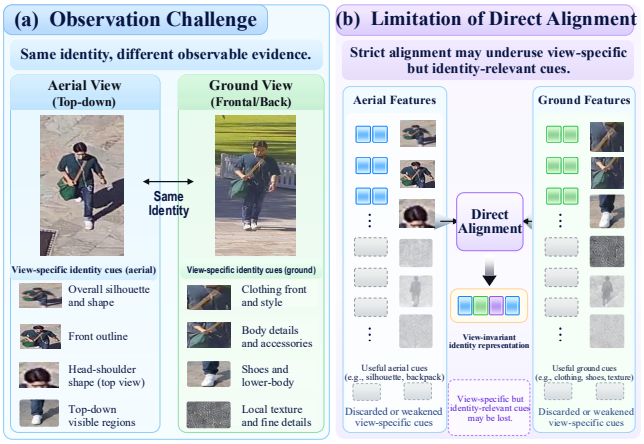

Figure 1: Motivation for PReVIEW. (a) Aerial and ground-based observations of the same identity exhibit distinct visual characteristics, as changes in viewpoint affect the appearance and visibility of identity cues. (b) Illustration of potential limitations inherent in direct feature alignment methods, highlighting that focusing solely on cross-view feature consistency may result in a failure to fully leverage viewpoint-specific yet identity-relevant cues (represented by faded elements in the diagram).

sent in single-view settings. Meanwhile, VR3D Ji et al. (2026) elevates identity representation to a unified 3D space, aiming to derive identity representations that are more robust and consistent across viewpoints by leveraging canonical 3D structures. Despite their varying designs, these methods primarily aim to bridge cross-view discrepancies by constructing representations that are more consistent or directly comparable between aerial and ground-level observations.

However, there are inherent differences in the observable identity cues between aerial and ground perspectives. This does not imply that cross-view matching must be achieved by eliminating such difference. Instead of forcing representation from the two views to align as closely as possible, we can preserve the identity information specific to each observational context and explicitly learn the predictive relationship that maintain identity consistency between them. This approach enables robust cross-view identity matching without the need to compress representation from different views into a highly uniform format.

As schematically illustrated in Figure 1(b), emphasizing direct feature alignment may underuse view-specific cues that remain informative about identity. Based on this observation, we propose a different perspective: rather than requiring aerial and ground representations to remain as aligned as possible, we allow them to retain features specific to their respective observation conditions while explicitly learning the associations between them that preserve identity information. The focus has shifted from pursuing perfect consistency between the two views to investigating whether there exists a stable, learnable, and generalizable predictive relationship between identity representations under varying observational conditions.

Therefore, we argue that the key lies not in devising stronger alignment constraints, but in explicitly learning an identity-conditioned predictive relationship across views. Given an identity's representation from one viewpoint, the model should be capable of predicting its latent representation from another. Unlike direct feature alignment, this predictive learning approach allows aerial and ground representations to retain their view-specific characteristics while enforcing a stable, identity-preserving mapping between them. In other words, the model is not required to force the two views to become identical; instead, it must understand how the representation of a specific identity transforms when the viewpoint changes.

To address this, we propose an identity-conditioned cross-view latent representation prediction mechanism (PReVIEW). During training, aerial and ground-level images of the same identity form cross-view prediction pairs, one view serves as the source observation to provide the current identity representation, while the other acts as the target observation to provide the latent target representa-

tion required for prediction. PReVIEW employs bidirectional global-to-regional latent prediction, where an encoder extracts the global representation of the source image. A predictor estimates the target view's representation based on the source representation; the source-side global context integrates information from the current observation, while the target-side regional representations preserve spatial distribution differences within the target observation. Since the prediction relies on the source image's global representation rather than source-side local features spatially aligned with the target regions, this design does not require a fixed one-to-one correspondence between cross-view regions. Regions serve solely to define latent prediction targets; they do not represent predefined body parts nor constitute an additional local retrieval branch. Meanwhile, the target branch utilizes a stop-gradient operation to provide a relatively stable supervision signal. This process operates simultaneously in both directions (aerial-to-ground and ground-to-aerial) and is jointly optimized with standard identity classification and metric learning tasks. Unlike generative methods, we do not need to reconstruct the target image or generate missing visual details. The prediction process takes place within the identity representation space, focusing on learning identity-related mappings across views. As the prediction branch functions only during training, with the standard re-identification pipeline maintained during inference, no additional computational overhead is incurred for prediction

Our contributions are summarized as follows:

- We introduce identity-conditioned cross-view predictability as a complementary training criterion for AG-ReID. Beyond using same-identity observations to impose identity consistency, we formulate cross-platform correspondence as a latent prediction task, learning predictive relations between different observations while retaining standard identity supervision.
- We propose PReVIEW, a bidirectional global-to-region latent prediction framework that does not require explicit view disentanglement or fixed cross-view part correspondence. The predictive objective jointly optimizes a lightweight predictor and the online identity encoder. The predictor and target encoder are used only during training, leaving the original retrieval architecture unchanged at inference.
- Experiments on CARGO show that PReVIEW improves a standard baseline across all four evaluation protocols. On LAGPeR, PReVIEW achieves the promising performance over previous methods. These results support our central insight that cross-view identity learning can benefit not only from representation consistency, but also from explicitly learning predictive relations between different observations of the same identity.

2 Related Work

2.1 Person Re-Identification

Person re-identification has made substantial progress in general Sun et al. (2019); Chen et al. (2019); He et al. (2021a) and occluded settings Wang et al. (2020); Tan et al. (2024); Ye et al. (2024). Ground-based approaches Li et al. (2014); He et al. (2023); Sun et al. (2018); Zhou et al. (2021) have benefited from benchmarks such as Market1501 and DukeMTMC-reID Zheng et al. (2015); Ristani et al. (2016). Conventional methods primarily address retrieval within homogeneous camera platforms, while recent unified frameworks accommodate diverse retrieval tasks through instructions, modalities, and interactive queries He et al. (2024); Li et al. (2024); Zuo et al. (2026); Niu et al. (2025). Nevertheless, aerial–ground matching introduces substantial viewpoint and visibility discrepancies that alter the available identity evidence, presenting additional challenges for cross-platform representation learning.

2.2 Aerial-Ground Person Re-Identification

representations through invariance learning, view-aware modeling, alignment, or complementary information integration. VDT and VIF suppress viewpoint bias through feature separation, orthogonal constraints, or rotation-aware regularization Zhang et al. (2024); Khalid et al. (2025), while SeCap and LATex exploit adaptive prompts and attribute-based textual guidance Wang et al. (2025); Zhang et al. (2025). ViSA and HiHR further preserve complementary view-dependent identity cues through

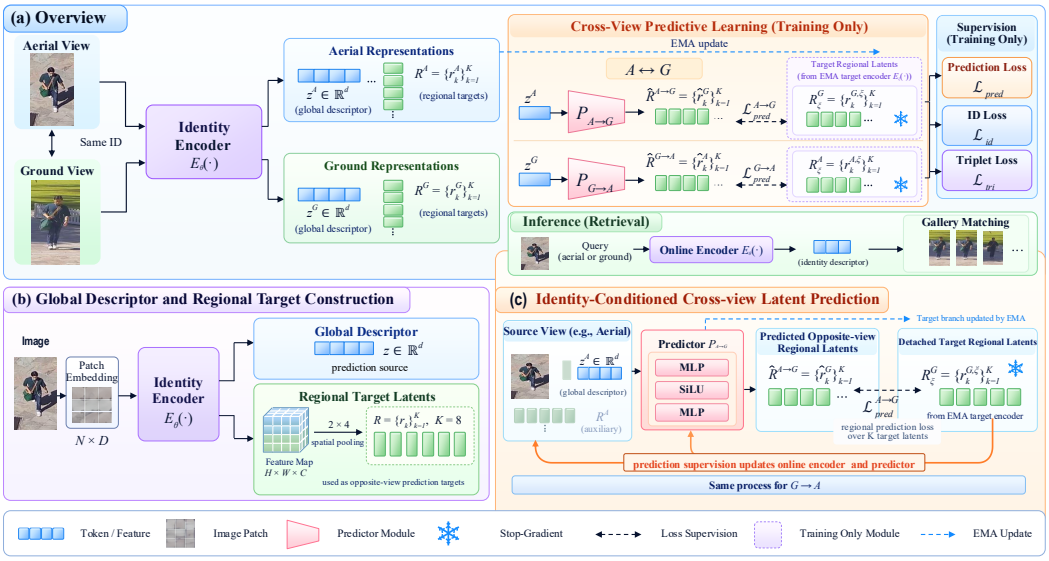

Figure 2: Overview of PReVIEW. (a) Same-identity aerial and ground observations form bidirectional latent prediction tasks, jointly optimized with identity classification and triplet supervision. (b) Global pooling produces the source representation z, while spatial pooling constructs regional representations. Prediction targets R_ξ are extracted by the EMA target encoder E_ξ. (c) The shared predictor, shown as $P_{A\to G}$ and $P_{G\to A}$ for its two conditional directions, predicts opposite-view regional latents from a source global representation. Prediction gradients update the online encoder and predictor, whereas targets are detached and E_ξ is updated from E_θ through EMA.

view-aware expert fusion and hierarchical representation learning Zhang et al. (2026); Yang & Zhang (2026). Other approaches address cross-view discrepancies through geometric and visibility-aware alignment Li et al. (2026) or generation of complementary view-specific features Wang et al. (2026). PReVIEW follows a different direction. Rather than explicitly decomposing, aligning, or completing cross-view features, we use another observation of the same identity as a latent prediction target, turning cross-platform identity correspondence into an additional learning signal for the retrieval encoder.

2.3 Predictive Representation Learning

Predictive representation learning has shown that latent prediction can provide effective supervision without reconstructing raw observations. Methods such as BYOL Grill et al. (2020) and SimSiam Chen & He (2021) learn from different augmentations of the same image instance, and I-JEPA Assran et al. (2023) predicts latent representations of target regions from their surrounding visual context. Inspired by this general principle, PReVIEW explores predictive learning in a different correspondence structure tailored to cross-view identity learning. Specifically, the source and target are distinct observations of the same identity captured by heterogeneous camera platforms, rather than two augmentations or spatial regions of a single image. Their relation is therefore defined by identity correspondence across viewing conditions. Based on this formulation, PReVIEW investigates whether cross-platform same-identity observations can serve as predictive targets for retrieval representation learning.

3 Method

Figure 2 presents PReVIEW, a framework that uses same-identity cross-platform prediction to improve identity representation learning without modifying the retrieval architecture at inference. Rather than imposing an additional constraint that directly aligns source and target features, PReVIEW learns a predictive mapping from the global context of one observation to the regional repre-

sentations of another observation of the same identity. The framework consists of an online identity encoder E_θ, an EMA target encoder E_ξ, and a shared platform-conditioned predictor. Identity classification and metric learning maintain the discriminability of the retrieval representation, while the predictive objective backpropagates through the predictor into the online encoder. The target encoder supplies detached regional targets. After training, only the online encoder and its original retrieval head are retained.

3.1 Global Descriptor and Regional Target Construction

Let $\mathcal{D} = \{(x_i, y_i, v_i)\}_{i=1}^{N}$ denote the training set, where x_i is a person image, $y_i \in \{1, \dots, C_{\mathrm{id}}\}$ is its identity label, and $v_i \in \{A, G\}$ indicates the aerial or ground platform. The online encoder E_θ is a pretrained Swin-B Liu et al. (2021) shared by both platforms. It maps an input image to a spatial feature map after the final layer normalization. Global average pooling and a batch-normalization neck then produce the pooled representation and retrieval descriptor:

$$\begin{aligned} H_i &= E_\theta(x_i) \in \mathbb{R}^{h \times w \times d}, \\ z_i &= \frac{1}{hw} \sum_{p=1}^{h} \sum_{q=1}^{w} H_i[p, q, :], \qquad b_i = \mathrm{BN}(z_i). \end{aligned} \tag{1}$$

For 256×128 inputs, the final feature map has $h = 8$, $w = 4$, and $d = 1024$. The predictor and triplet objective operate on z_i, whereas identity classification and retrieval use the BNNeck feature b_i. Superscripts such as z_i^A and z_i^G indicate the observation platform rather than different encoders. As summarized in Figure 2(b), the same spatial representation also supports regional feature construction. However, all regional targets used for predictive supervision are extracted by the target encoder, not taken directly from the online branch.

The target encoder E_ξ has the same architecture as E_θ and is initialized with its parameters. It receives no gradient-based parameter updates. Immediately before extracting target features at training step t, we update its parameters as

$$\xi_t = m\xi_{t-1} + (1 - m)\theta_t, \tag{2}$$

where θ_t denotes the online parameters before the current optimizer step. The target encoder remains in evaluation mode, and its non-parameter buffers are copied from the online encoder. For a target image x_j from platform $b = v_j$, let $\bar{H}_j = E_\xi(x_j)$. We partition this feature map into a fixed 2×4 grid and average the tokens within each region Ω_k:

$$\begin{aligned} r_{j,k}^{b,\xi} &= \mathrm{sg}\left(\frac{1}{|\Omega_k|} \sum_{(p,q) \in \Omega_k} \bar{H}_j[p, q, :] \right) \in \mathbb{R}^d, \\ R_{\xi,j}^{b} &= \{r_{j,k}^{b,\xi}\}_{k=1}^{K}, \qquad K = 8. \end{aligned} \tag{3}$$

Here, sg denotes stop-gradient. Each region contains four spatial tokens on the 8×4 feature map. These regions define locations within the target observation; they do not denote predefined body parts or assume corresponding locations in the source image. The source global representation provides the prediction context, while the target regions retain spatially differentiated supervision. Source-side regional features are not additional visual inputs to the predictor.

3.2 Identity-Conditioned Cross-view Latent Prediction

We construct prediction pairs from images with the same identity but different platform labels. Pairing is performed within each device's local mini-batch $\mathcal{B}$. For identity y, let A_y and G_y be the ordered lists of its aerial and ground indices. When both lists are nonempty, the aerial source at zero-based position s is paired with the ground image at position $s \bmod |G_y|$; the reverse direction is constructed analogously. This produces the directed pair sets $\mathcal{P}_{A \to G}$ and $\mathcal{P}_{G \to A}$. Every eligible image is used once as a source in its corresponding direction, while a target may be reused when the platform counts differ. Identities observed on only one platform contribute to the identity objectives but do not form prediction pairs. Identity conditioning therefore determines which observations are paired; the identity label itself is not an input to the predictor.

For a pair $(i, j) \in \mathcal{P}_{a\to b}$, the predictor estimates the regional targets of platform b from the source global representation z_i^a. A learnable region query $q_k \in \mathbb{R}^{128}$ specifies the target region, and a shared embedding function $e(v) \in \mathbb{R}^{64}$ encodes the source and target platforms:

$$\begin{aligned} \widehat{r}_{i,k}^{\,b} &= P_{a\to b}(z_i^a, q_k) := g_\phi([z_i^a; q_k; e(a); e(b)]) \,, \\ \widehat{R}_i^{a\to b} &= \{\widehat{r}_{i,k}^{\,b}\}_{k=1}^{K}, \end{aligned} \tag{4}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation. The shared predictor g_ϕ is a two-layer MLP with a SiLU activation, a hidden dimension of 1024, and an output dimension of 1024. The notation $P_{A\to G}$ and $P_{G\to A}$ in Figure 2 denotes two platform-conditioned applications of this same predictor, not separate parameter sets. Region queries and platform embeddings are also shared across directions. The target image is used only to construct the supervisory regional representations; its visual features are not predictor inputs. Platform embeddings condition the auxiliary prediction task and are not supplied to the retrieval encoder.

For a direction $a \to b$ with a nonempty pair set, we define the regional prediction objective as

$$\mathcal{L}_{\text{pred}}^{a\to b} = \frac{1}{K|\mathcal{P}_{a\to b}|} \sum_{(i,j)\in\mathcal{P}_{a\to b}} \sum_{k=1}^{K} \left\| \widehat{r}_{i,k}^{\,b} - r_{j,k}^{b,\xi} \right\|_2^2 . \tag{5}$$

The squared error is summed over feature coordinates and averaged over regions and prediction pairs. Neither predicted nor target regional features are ℓ_2-normalized for this objective. Prediction errors are accumulated in float32 under mixed-precision training. Let $\mathcal{V} = \{a \to b : |\mathcal{P}_{a\to b}| > 0\}$ denote the available prediction directions. The bidirectional objective is

$$\mathcal{L}_{\text{pred}} = \begin{cases} \dfrac{1}{|\mathcal{V}|} \displaystyle\sum_{a\to b\in\mathcal{V}} \mathcal{L}_{\text{pred}}^{a\to b}, & |\mathcal{V}| > 0, \\ 0, & |\mathcal{V}| = 0. \end{cases} \tag{6}$$

Thus, the two directions receive equal weight when both are present, irrespective of their respective pair counts. Prediction gradients update the shared predictor, region queries, platform embeddings, and online source encoder, while the regional targets remain detached. As illustrated in Figure 2(c), the objective supervises a mapping from global identity context to opposite-view regional representations, rather than requiring the source global feature to directly coincide with the target features.

3.3 Joint Optimization and Retrieval

Predictive learning supplements the standard identity classification and metric objectives. Let $B = |\mathcal{B}|$ be the local batch size and w_c the classifier weight for identity c. Identity classification is applied to the BNNeck feature b_i. For metric learning, we perform batch-hard mining on the unnormalized pooled features z_i within the same local batch. The hardest positive and negative distances are $d_i^+ = \max_{j\in\mathcal{B},\, y_j=y_i,\, j\neq i} \|z_i - z_j\|_2$ and $d_i^- = \min_{j\in\mathcal{B},\, y_j\neq y_i} \|z_i - z_j\|_2$, respectively. The two identity objectives are

$$\mathcal{L}_{\text{id}} = -\frac{1}{B} \sum_{i\in\mathcal{B}} \log \frac{\exp(w_{y_i}^\top b_i)}{\sum_{c=1}^{C_{\text{id}}} \exp(w_c^\top b_i)}, \tag{7}$$

$$\mathcal{L}_{\text{tri}} = \frac{1}{B} \sum_{i\in\mathcal{B}} \log\big(1 + \exp(d_i^+ - d_i^-)\big) \,. \tag{8}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}} + \lambda_{\text{pred}} \mathcal{L}_{\text{pred}}, \qquad \lambda_{\text{pred}} = 0.2. \tag{9}$$

The identity objectives train discriminative and comparable retrieval representations, while predictive supervision additionally requires the pooled source representation to support the estimation of another same-identity observation. Because the predictive loss acts on the online encoder through z_i, its influence is not restricted to the auxiliary predictor. This joint optimization does not introduce an independent view-feature branch, a view classification loss, or explicit orthogonal feature decomposition. The remaining optimization settings are provided in Section 4.2.

At inference, an image is represented by the online encoder's d-dimensional BNNeck feature $b(x) = \mathrm{BN}(\mathrm{GAP}(E_\theta(x)))$, with $d = 1024$ in our implementation. Given query and gallery images x_q and x_g, retrieval uses cosine distance:

$$d(x_q, x_g) = 1 - \frac{b(x_q)^\top b(x_g)}{\|b(x_q)\|_2 \, \|b(x_g)\|_2}. \tag{10}$$

The identity classifier, target encoder, predictor, region queries, and platform embeddings are discarded for retrieval. No target observation or platform condition is needed to extract a query or gallery descriptor. PReVIEW therefore adds predictive computation during training while preserving the inference architecture and descriptor dimension of the standard Swin-B baseline.

4 Experiments

We evaluate PReVIEW on five aerial–ground person re-identification benchmarks. This section primarily presents the key comparative results for CARGO and LAGPeR, analyzes the proposed PreVIEW. Complete evaluations on AG-ReID, AG-ReID.v2, CP2108 and G2APS-ReID are provided in Appendix A.1, with their principal findings summarized below.

4.1 Datasets and Evaluation Protocol

The experiments on five aerial-ground person re-identification datasets, CARGO Zhang et al. (2024), AGReID Nguyen et al. (2023), AGReIDv2 Nguyen et al. (2024), G2APS-ReID Wang et al. (2025), CP2108 Zheng et al. (2026) and LAGPeR Wang et al. (2025), ensuring no overlap between training and testing identities. The perspectives are denoted as A (Aerial), G (Ground), C (Fixed CCTV), and W (Wearable).

CARGO: An aerial-ground cross-view person ReID dataset with 5 aerial and 8 ground cameras, supporting both cross-view and same-platform person matching.

LAGPeR: A large-scale real-world aerial-ground ReID dataset with 4,231 identities and 63,841 images across 7 scenes and 21 cameras, covering aerial, ground oblique, and ground frontal views with complex scene, camera, and viewpoint variations. Complete introduction on AG-ReID, AG-ReID.v2, and G2APS-ReID are provided in Appendix A.1

4.2 Implementation Details

Experiments were conducted using PyTorch 1.12.1 on an NVIDIA RTX 4090. A pretrained Swin-B Liu et al. (2021) served as the backbone, with an input size of 256×128 and a global batch size of 128, comprising 32 identities with four images per identity. The model was trained for 120 epochs using SGD with a momentum of 0.9 and a base learning rate of 8×10^{-3}. We used linear warm-up over the first 2,000 iterations with an initial factor of 0.01, together with a cosine learning-rate schedule with a minimum learning rate of 1.6×10^{-6}. The target encoder was updated through exponential moving average (EMA) with a momentum coefficient of 0.999, and the target features were detached from backpropagation. Target spatial features were partitioned into a 2×4 grid, yielding eight regional representations. The predictor was a two-layer MLP with a hidden dimension of 1024, while the dimensions of the region queries and platform embeddings were 128 and 64, respectively. The prediction loss $\mathcal{L}_{\mathrm{pred}}$ was weighted by $\lambda = 0.2$, whereas the plain Swin-B baseline was trained only with identity classification and triplet losses. Neither the predictor nor the target encoder was used during inference.

4.3 Comparison with State-of-the-art Methods

Although PReVIEW does not achieve the best performance under the challenging CARGO $A \leftrightarrow G$ protocol, it shows strong overall competitiveness across the evaluated retrieval settings. Under the ALL protocol, PReVIEW achieves 76.28% R1 and 68.46% mAP, outperforming the previous best results by 1.92 and 0.94 percentage points, respectively. It also obtains the highest $A \leftrightarrow A$ mAP of 76.12%, exceeding the previous best result by 5.61 points. Compared with the matched Swin-B baseline, PReVIEW improves R1/mAP by 11.87/10.08 points under $A \leftrightarrow G$ and 12.01/6.49 points under ALL.

Table 1: Comparison with representative methods on CARGO. Results of previous methods are quoted from their original publications under the reported CARGO settings. R1 and mAP are reported as percentages. "–" denotes results not reported in the original paper; SVPR-ReID reports only mAP under ALL and $A \leftrightarrow G$. The best and second-best distinct results are highlighted in bold and underlined, respectively, including ties. Δ denotes percentage-point gains over the matched Swin-B baseline.

Method	Pub.	$A \leftrightarrow G$		ALL		$G \leftrightarrow G$		$A \leftrightarrow A$	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
VDT Zhang et al. (2024)	CVPR'24	48.12	42.76	64.10	55.20	82.14	71.59	**82.50**	66.83
DTST Wang & Pishgar (2025)	ICME'25	50.63	43.39	64.42	55.73	78.57	72.40	80.00	63.31
VIF-AGReID Khalid et al. (2025)	ICCV'25	51.25	44.55	65.71	57.46	83.93	74.19	**82.50**	66.98
SeCap Wang et al. (2025)	CVPR'25	69.43	58.94	68.59	60.19	86.61	75.42	80.00	68.08
GSAlign Li et al. (2026)	NeurIPS'25	64.89	61.55	65.06	57.95	83.04	73.86	80.00	65.55
SD-ReID Wang et al. (2026)	TIP'26	53.12	46.44	65.06	57.47	81.25	74.08	**82.50**	67.70
ViSA Zhang et al. (2026)	CVPR'26	71.28	69.00	70.51	65.46	**88.39**	**83.90**	**82.50**	67.78
SVPR-ReID Zheng et al. (2026)	AAAI'26	-	64.05	-	66.00	-	-	-	-
VR3D Ji et al. (2026)	arXiv'26	59.38	54.57	71.15	64.48	84.82	78.74	**82.50**	70.51
HiHR Yang & Zhang (2026)	ECCV'26	**75.53**	**70.53**	74.36	67.52	87.50	80.93	75.00	69.60
Swin-B (baseline)	–	56.88	51.17	64.27	61.97	82.14	75.23	77.50	66.44
PReVIEW	Ours	68.75	61.25	**76.28**	**68.46**	85.71	79.30	80.00	**76.12**
Δ over Swin-B	–	+11.87	+10.08	+12.01	+6.49	+3.57	+4.07	+2.50	+9.68

Table 2: Comparison with representative methods on LAGPeR. R1 and mAP are reported as percentages. The best and second-best distinct results are highlighted in bold and underlined, respectively. Δ denotes percentage-point gains over the matched Swin-B baseline.

Method	Pub.	$A \rightarrow G$		$G \rightarrow A$		$G \rightarrow (A+G)$	
		R1	mAP	R1	mAP	R1	mAP
ViT Dosovitskiy (2020)	ICLR'21	38.67	27.25	32.04	30.69	18.88	15.31
TransReID He et al. (2021b)	ICCV'21	38.80	28.80	33.00	32.10	22.90	18.80
VDT Zhang et al. (2024)	CVPR'24	40.15	28.97	33.55	31.98	19.50	16.45
SeCap Wang et al. (2025)	CVPR'25	41.79	30.37	35.26	33.42	24.39	19.24
ViSA Zhang et al. (2026)	CVPR'26	41.37	30.42	35.06	33.59	25.61	20.31
Swin-B (baseline)	–	32.44	22.57	27.22	25.71	15.43	12.53
PReVIEW	–	**48.46**	**36.30**	**40.64**	**39.57**	**31.71**	**25.93**
Δ over Swin-B	–	+16.02	+13.73	+13.42	+13.86	+16.28	+13.40

On LAGPeR, PReVIEW ranks first on all metrics in Table 2. These results complement the improvements over the matched Swin-B baseline, demonstrating that the benefit of predictive supervision translates into competitive absolute retrieval performance rather than only relative gains over the baseline. Crucially, PReVIEW achieves these results while retaining the original Swin-B encoder and retrieval head at inference, without explicit view-feature decomposition or augmentation of the retrieval descriptor with predicted regional features. Instead, regional representations of another same-identity observation serve as training targets, and their prediction errors jointly optimize the online encoder with the identity objectives.

4.4 Ablation Studies

Effect of prediction direction We compare aerial-to-ground, ground-to-aerial, and bidirectional prediction in Table 3. Both unidirectional variants improve the baseline, while bidirectional prediction yields the strongest overall performance, achieving 76.28/68.46 R1/mAP under ALL and 68.75/61.25 under $A \leftrightarrow G$. Compared with the best unidirectional result for each metric, the bidirectional setting improves ALL by 4.49/3.62 points and $A \leftrightarrow G$ by 3.75/5.05 points. It also performs best under $A \leftrightarrow A$, while under $G \leftrightarrow G$ it ties the best R1 but remains below the $G \rightarrow A$ variant in mAP. Overall, these results support jointly modeling predictive relations in both directions rather than relying on a single cross-view prediction task.

Table 3: Effect of prediction direction on CARGO. Columns specify the training-time prediction configuration, while rows denote retrieval protocols. The baseline uses identity classification and triplet losses without predictive supervision. Each entry reports R1/mAP (%). The best and second-best distinct values for each metric within each row are highlighted in bold and underlined, respectively, including ties.

Protocol	Baseline	$A \to G$ only	$G \to A$ only	Bidirectional
ALL	64.27 / 61.97	71.79 / 64.66	70.19 / 64.84	**76.28 / 68.46**
$A \leftrightarrow G$	56.88 / 51.17	65.00 / 56.20	57.50 / 53.42	**68.75 / 61.25**
$A \leftrightarrow A$	77.50 / 66.44	75.00 / 70.04	77.50 / 68.06	**80.00 / 76.12**
$G \leftrightarrow G$	82.14 / 75.23	81.25 / 77.32	**85.71 / 82.07**	**85.71** / 79.30

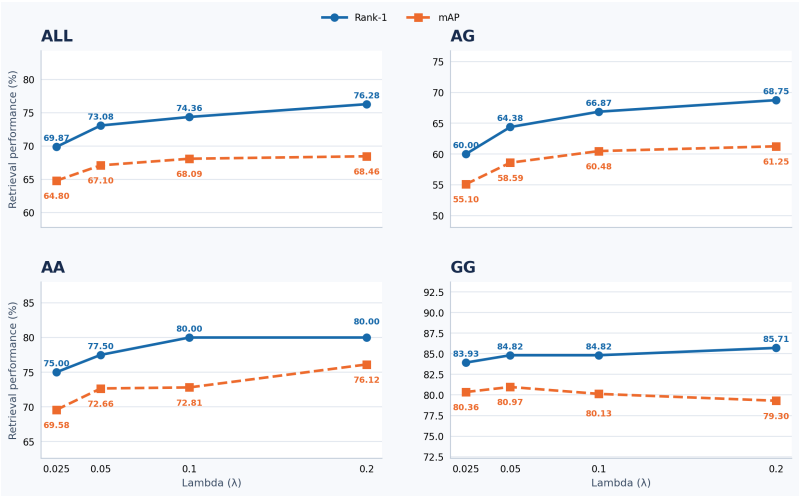

Figure 3: Effect of the prediction-loss weight λ_{pred} on CARGO under the ALL, $A \leftrightarrow G$, $A \leftrightarrow A$, and $G \leftrightarrow G$ protocols. R1 and mAP are reported as percentages.

5 Visual Analysis

Parameter visualization Figure 3 shows that increasing λ_{pred} steadily improves the ALL and $A \leftrightarrow G$ protocols, reaching 76.28/68.46 and 68.75/61.25 R1/mAP at $\lambda = 0.2$. Performance on $A \leftrightarrow A$ also improves. This shows that stronger predictive supervision is particularly beneficial for cross-view retrieval, and we set $\lambda_{\text{pred}} = 0.2$ by default. Additional visualization results are provided in Appendix A.3.

6 Conclusion

We presented PReVIEW, a cross-view predictive learning framework for aerial–ground person re-identification. Instead of relying solely on representation consistency across views, PReVIEW treats same-identity observations from different camera platforms as predictive counterparts and learns bidirectional latent relations between them. The resulting predictive supervision directly optimizes the identity encoder, while the predictor and target encoder are removed after training.

Experiments across multiple aerial–ground ReID benchmarks demonstrate consistent improvements over the matched baseline, together with strong performance on CARGO and LAGPeR. These results support our central insight that cross-view identity correspondence can serve not only as a matching constraint, but also as a useful predictive learning signal for retrieval representation learning.

AI USE STATEMENT

In this work, we used generative AI tools only as a writing assistant for language polishing, including improvements to grammar, fluency, clarity, and conciseness. The manuscript itself was authored by the researchers; the research ideas, problem formulation, methodology, experimental design, and result analysis were developed independently by the authors rather than generated by AI. All AI-assisted language edits were manually reviewed and revised to ensure technical accuracy, originality, and consistency with the actual experiments.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023. 4

Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8350–8360. IEEE, 2019. 3

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021. 4

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4

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*, pp. 9664–9667, 2023. 3

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 14993–15002. IEEE, 2021a. 3

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 14993–15002. IEEE, 2021b. 8

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17521–17531. IEEE, 2024. 3

Chao Ji, Shiyu Xuan, and Zechao Li. Vr3d: View-robust 3d representation learning for aerial-ground person re-identification. *arXiv preprint arXiv:2608.02598*, 2026. 2, 8, 13

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*, pp. 9749–9758, 2025. 3, 8, 13

He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17459–17469. IEEE, 2024. 3

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. *Advances in Neural Information Processing Systems*, 38:93785–93805, 2026. 1, 4, 8, 13

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159, 2014. 3

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021. 5, 7

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2585–2590. IEEE, 2023. 7, 13

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. 7, 13

Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24245–24254. IEEE, 2025. 3

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016. 3

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *European conference on computer vision*, pp. 501–518. Springer, 2018. 3

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. 3

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, pp. 5070–5078, 2024. 3

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6448–6457. IEEE, 2020. 3

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22119–22128. IEEE, 2025. 3, 7, 8, 13

Yuhai Wang and Maryam Pishgar. Dynamic token selective transformer for aerial-ground person re-identification. In *IEEE International Conference on Multimedia and Expo*, 2025. 8, 13

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: view-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 2026. 1, 4, 8, 13

Qiwei Yang and Pingping Zhang. Hihr: Hierarchical hyperbolic representation for aerial-ground person re-identification. *arXiv preprint arXiv:2607.09186*, 2026. 4, 8, 13

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, pp. 6684–6692, 2024. 3

Pingping Zhang, Xiang Hu, Yuhao Wang, and Huchuan Lu. Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. *arXiv preprint arXiv:2503.23722*, 2025. 3

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*, pp. 22000–22009, 2024. 1, 3, 7, 8, 13

Quan Zhang, Zeqiang Cai, Peiming Zhao, Jingze Wu, Cailun Wu, Hongbo Chen, and Jianhuang Lai. View-aware semantic alignment for aerial-ground person re-identification. *arXiv preprint arXiv:2605.18192*, 2026. 1, 4, 8, 13

A. Zheng, H. Xie, X. Wan, Z. Wang, S. Li, J. Tang, and B. Luo. Semantic-driven visual progressive refinement for aerial-ground person ReID: A challenging large-scale benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13360–13368, 2026. 7, 8, 12, 13, 14, 15

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015. 3

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. 3

Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. Reid5o: Achieving omni multi-modal person re-identification in a single model. *Advances in Neural Information Processing Systems*, 38:53401–53424, 2026. 3

A Appendix

A.1 Additional Evaluation

This section evaluates whether the benefits of same-identity cross-platform predictive supervision extend beyond our primary CARGO benchmark. We consider AG-ReID, AG-ReID.v2, and G2APS-ReID, which provide complementary camera configurations and retrieval settings. Models are trained separately on each dataset.

AGReID: Contains 388 identities and 21,983 drone/ground-view images, with key challenges in shooting angle, person scale, and visible body region disparities between aerial and ground views.

AGReIDv2: Expands AGReID to 1,615 identities and 100,502 images from drones, fixed surveillance cameras, and wearable cameras, with 15 attribute categories for multi-view and attribute-assisted ReID.

G2APS-ReID: Reconstructed from G2APS by extracting identity-annotated person crops, it is an aerial-ground cross-view ReID dataset with large viewpoint, scale, and background differences for complex cross-view matching.

CP2108 Zheng et al. (2026): A large-scale real-world aerial–ground ReID benchmark containing 142,817 images of 2,108 identities captured by 17 UAVs and 22 ground cameras across diverse scenes and time conditions. Across these experiments, PReVIEW retains the standard encoder and retrieval head. Its additional global-to-region prediction task is used only during training and does not require an independent view-feature branch, orthogonal disentanglement, or predicted regional features during retrieval.

Table 4: Comparison with representative methods on AG-ReID. R1 and mAP are reported as percentages. The best and second-best distinct values are highlighted in bold and underlined, respectively, including ties. Δ denotes percentage-point gains over our Swin-B baseline.

Method	Pub.	$A \to G$		$G \to A$	
		R1	mAP	R1	mAP
Explain Nguyen et al. (2023)	ICME'23	81.47	72.61	82.85	73.39
VDT Zhang et al. (2024)	CVPR'24	82.91	74.44	86.59	78.57
DTST Wang & Pishgar (2025)	ICME'25	83.48	74.51	84.72	76.05
VIF-AGReID Khalid et al. (2025)	ICCV'25	83.75	75.22	87.32	79.19
SeCap Wang et al. (2025)	CVPR'25	84.03	76.16	87.01	78.34
GSAlign Li et al. (2026)	NeurIPS'25	<u>86.74</u>	**84.00**	87.94	**87.17**
SD-ReID Wang et al. (2026)	TIP'26	85.16	75.40	85.97	77.02
SVPR-ReID Zheng et al. (2026)	AAAI'26	85.34	77.85	87.32	80.55
HiHR Yang & Zhang (2026)	ECCV'26	86.06	79.21	87.94	81.28
VR3D Ji et al. (2026)	arXiv'26	**87.39**	<u>79.95</u>	**89.60**	<u>82.87</u>
Swin-B (baseline)	–	77.59	67.08	78.27	70.15
PReVIEW	Ours	86.50	79.25	<u>88.73</u>	82.80
Δ over Swin-B	–	+8.91	+12.17	+10.46	+12.65

Table 5: Comparison with representative methods on AG-ReID.v2. A, C, and W denote aerial, CCTV, and wearable cameras, respectively. R1 and mAP are reported as percentages. The best and second-best distinct values are highlighted in bold and underlined, respectively, including ties. Δ denotes percentage-point gains over our Swin-B baseline.

Method	Pub.	$A \to C$		$A \to W$		$C \to A$		$W \to A$	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
Explain Nguyen et al. (2023)	ICME'23	87.70	79.00	**93.67**	83.14	87.35	78.24	87.73	79.08
VDT Zhang et al. (2024)	CVPR'24	86.46	79.13	90.00	82.21	86.14	78.12	85.26	78.52
V2E Nguyen et al. (2024)	TIFS'24	88.77	80.72	<u>93.62</u>	84.85	87.86	78.51	88.61	80.11
SeCap Wang et al. (2025)	CVPR'25	88.12	80.84	91.44	84.01	88.24	79.99	87.56	80.15
GSAlign Li et al. (2026)	NeurIPS'25	87.86	81.38	90.63	83.98	88.02	81.05	87.31	80.90
SD-ReID Wang et al. (2026)	TIP'26	87.04	80.61	90.86	84.06	86.74	79.24	86.79	80.12
ViSA Zhang et al. (2026)	CVPR'26	<u>89.43</u>	83.61	91.63	85.99	<u>88.57</u>	82.32	89.23	83.10
SVPR-ReID Zheng et al. (2026)	AAAI'26	–	<u>84.35</u>	–	–	–	–	–	80.64
HiHR Yang & Zhang (2026)	ECCV'26	88.84	84.04	91.40	<u>87.57</u>	<u>88.57</u>	<u>83.21</u>	<u>89.53</u>	<u>84.02</u>
VR3D Ji et al. (2026)	arXiv'26	**89.98**	**84.54**	93.57	**87.96**	**90.56**	**84.39**	**89.70**	**84.17**
Swin-B (baseline)	–	83.70	76.16	87.19	79.90	83.93	75.30	84.74	76.70
PReVIEW	Ours	87.63	81.07	90.46	83.77	86.81	79.41	86.98	81.00
Δ over Swin-B	–	+3.93	+4.91	+3.27	+3.87	+2.88	+4.11	+2.24	+4.30

AG-ReID Table 4 reports results in both aerial–ground retrieval directions. PReVIEW achieves 86.50% R1 and 79.25% mAP for $A \to G$, and 88.73% R1 and 82.80% mAP for $G \to A$. Compared with the Swin-B baseline, the R1 gains are 8.91 and 10.46 percentage points, while the mAP gains are 12.17 and 12.65 points, respectively.

The improvements in both directions show that the benefit of predictive training extends to real aerial–ground observations and is not confined to a single query platform. PReVIEW also exceeds the reported VDT.

AG-ReID.v2 Table 5 evaluates PReVIEW across aerial, CCTV, and wearable observations. PReVIEW improves both R1 and mAP over the Swin-B baseline in all four protocols. The R1 gains range from 2.24 to 3.93 percentage points, and the mAP gains range from 3.87 to 4.91 points. Each exceeds the corresponding reported VDT mAP.

The gains across both CCTV and wearable settings extend the evidence beyond a single type of ground observation. Although these camera configurations expose different visual evidence, the same predictive-learning formulation improves the corresponding identity representations. No ad-

Table 6: Additional results on G2APS-ReID. The + notation follows the evaluated gallery settings. R1 and mAP are reported as percentages. Bold indicates the better result within each column. Δ denotes percentage-point gains over the Swin-B baseline.

Method	$A \to G^+$		$G \to A^+$	
	R1	mAP	R1	mAP
Swin-B (baseline)	66.35	49.63	62.63	48.34
PReVIEW	**68.19**	**53.69**	**64.54**	**52.10**
Δ over Swin-B	+1.84	+4.06	+1.91	+3.76



Figure 4: Representative Rank-1 retrieval results on CARGO under the $A \leftrightarrow G$ protocol. For each query, the middle and right images show the Rank-1 results of Swin-B and PReVIEW, respectively. Green and red boxes indicate correct and incorrect matches.

ditional camera-specific retrieval branch is introduced: platform conditions are used within the training-time predictor, while inference retains the standard encoder and retrieval head.

G2APS-ReID We provide an additional within-backbone evaluation on G2APS-ReID under the $A \to G^+$ and $G \to A^+$ settings used in our experiments. The Swin-B baseline and PReVIEW are evaluated under the same protocol variant in each direction. As in the main experiments, PReVIEW adds cross-platform predictive supervision during training and retains only the online encoder and retrieval head at inference.

A.2 Ablation Studies

As shown in Table 6, PReVIEW achieves 68.19% R1 and 53.69% mAP for $A \to G^+$, improving the baseline by 1.84 and 4.06 percentage points. For $G \to A^+$, it achieves 64.54% R1 and 52.10% mAP, with gains of 1.91 and 3.76 points, respectively. These positive changes extend the within-backbone evidence to another evaluation setting. Including G2APS-ReID, PReVIEW improves both R1 and mAP in all 15 reported protocols across the five separately trained datasets.

CP2108 We further evaluate PReVIEW on CP2108 Zheng et al. (2026), a recent large-scale real-world aerial–ground ReID benchmark. Following the original evaluation protocol, we report results under ALL, $A \to G$, and $G \to A$. As shown in Table 7, PReVIEW consistently outperforms SVPR-ReID across all three settings. In particular, PReVIEW improves R1/mAP by 3.84/2.77 points under

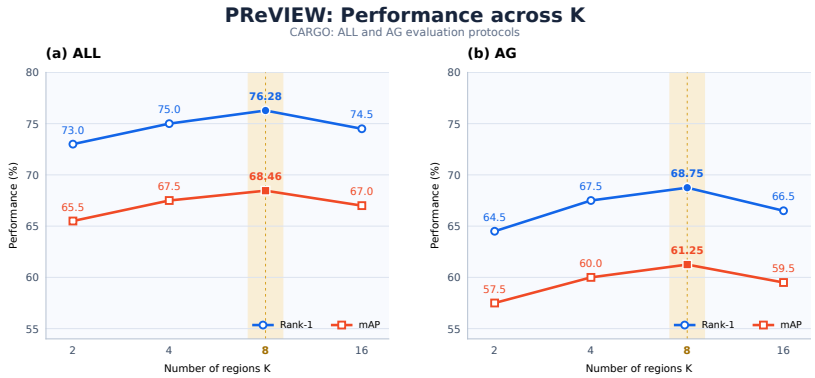

Figure 5: Effect of the number of regional prediction targets K on CARGO under the ALL and $A \leftrightarrow G$ protocols. R1 and mAP are reported as percentages.

Table 7: Comparison on CP2108 under the original evaluation protocols. R1 and mAP are reported as percentages. Δ denotes percentage-point gains over SVPR-ReID.

Method	ALL		$A \to G$		$G \to A$	
	R1	mAP	R1	mAP	R1	mAP
SVPR-ReID Zheng et al. (2026)	59.23	42.40	64.25	47.72	63.76	47.04
PReVIEW	**63.07**	**45.17**	**70.41**	**51.67**	**70.09**	**50.22**
Δ	+3.84	+2.77	+6.16	+3.95	+6.33	+3.18

ALL, 6.16/3.95 points under $A \to G$, and 6.33/3.18 points under $G \to A$. These results provide further evidence that same-identity cross-platform predictive supervision remains effective on challenging real-world aerial–ground observations.

A.3 Visual Analysis

Retrieval Visualization Figure 4 presents representative Rank-1 retrieval results on CARGO under the $A \leftrightarrow G$ protocol. For the same queries, the Swin-B baseline often ranks visually similar but incorrect identities first, particularly under large viewpoint, pose, and appearance variations. In contrast, PReVIEW retrieves the correct cross-view identity at Rank-1 in these challenging cases. These examples qualitatively complement the quantitative gains in Rank-1 accuracy, demonstrating that cross-view prediction supervision optimizes the ranking of preserved identity representations—an advantage that persists even after the prediction branch is removed during inference.

Effect of the number of regional targets. We vary the number of target regions K in Figure 5. Performance improves consistently from $K = 2$ to $K = 8$, reaching 76.28/68.46 R1/mAP under ALL and 68.75/61.25 under $A \leftrightarrow G$. Further increasing K to 16 leads to a moderate decline, suggesting that overly fine partitioning may reduce the effectiveness of regional prediction targets. We therefore set $K = 8$ by default.