VIGA: View-Conditioned and Identity-Guided Adaptation for Aerial-Ground Person Re-Identification

Xu Zhang[1,2,3], Yining Feng[1], and Zuyu Zhang[1,2,3⋆]

[1] School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China
zhangx@cqupt.edu.cn, s240233016@stu.cqupt.edu.cn, zhangzuyu@cqupt.edu.cn
[2] Academy of Advanced Interdisciplinary Studies, Chongqing University of Posts and Telecommunications, Chongqing, China
[3] Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing, China

Abstract. Aerial-ground person re-identification (AG-ReID) is significantly challenged by heterogeneous geometric regimes and severe foreground-background imbalance. Traditional shared-backbone architectures suffer from transformation conflicts when modeling disparate aerial and ground views, while residual connections in transformers often lead to identity embedding contamination from persistent background activations. To address these issues, we propose View-conditioned and Identity-Guided Adaptation (VIGA), a unified framework that mitigates AG-ReID-specific aerial-ground discrepancies at both the parameter and activation levels. VIGA incorporates View-Conditioned Low-Rank Adaptation to dynamically modulate feed-forward weights using view-aware low-rank residuals, thereby alleviating view-dependent geometric mapping conflicts without substantial parameter overhead. Simultaneously, we introduce Identity-Guided Sparsification, which employs a soft masking mechanism guided by identity semantics to suppress irrelevant background noise while preserving the 2D spatial topology. Extensive evaluations on the LAGPeR and AG-ReID.v2 datasets demonstrate that VIGA achieves state-of-the-art performance, particularly in challenging cross-platform retrieval scenarios, by effectively learning robust, view-invariant representations.

Keywords: Aerial-Ground Re-ID · View-Conditioned LoRA · Identity-Guided Sparsification · View-Dependent Adaptation · Dynamic Network

1 Introduction

Person re-identification (Re-ID) aims to retrieve images of the same individual across disjoint camera views and has achieved substantial progress with convolutional networks [16, 22] and vision transformers [3, 6]. With the rapid deployment of unmanned aerial vehicles (UAVs), a more challenging scenario has

⋆ Corresponding author.

emerged: Aerial-Ground Re-ID (AG-ReID) [12, 13]. Unlike conventional ground-to-ground settings, AG-ReID requires matching identities between aerial top-down imagery and ground-level surveillance cameras, introducing severe foreground–background imbalance and extreme geometric discrepancies.

A central difficulty arises from token-space imbalance in aerial imagery. Due to the wide field of view of UAVs, pedestrians typically occupy only a small portion of the image and are surrounded by extensive background clutter [19, 21]. Transformer-based models rely on attention mechanisms to reweight token importance, yet softmax normalization suppresses irrelevant responses only relatively. Low-weight background activations may persist through residual connections and accumulate across layers, progressively contaminating identity embeddings. Existing approaches attempt to mitigate this issue through refined attention mechanisms [19] or dynamic token pruning [14]. However, relative reweighting does not alter activation magnitude, and hard pruning often disrupts spatial continuity and optimization stability. These limitations suggest that AG-ReID requires explicit control over activation magnitude to prevent cumulative noise propagation in token space.

Beyond token-level interference, AG-ReID also suffers from structural mismatch in transformation space. Aerial projection introduces anisotropic deformation and scale compression, causing identity cues to undergo view-dependent geometric shifts [13, 20]. When a shared backbone is optimized across aerial and ground domains, its transformation operators are implicitly required to approximate heterogeneous geometric mappings. Such shared optimization tends to produce a compromised transformation that is suboptimal for both views, weakening discriminative consistency in the embedding space. Recent studies explore feature decoupling, such as View-Decoupled Transformer (VDT) [20], prompt-based view modeling [18], or cross-modal alignment [8, 9]. Nevertheless, most methods adapt representations while keeping backbone parameters static, leaving geometric conflicts unresolved at the operator level.

To address these coupled challenges, we propose a unified framework, named VIGA, composed of two complementary components. First, we introduce Identity-Guided Sparsification (IGS), which estimates semantic relevance between global identity embeddings and local patches and attenuates irrelevant activations to near-zero magnitude while preserving spatial structure. By reducing background responses at the activation level, IGS limits cumulative noise amplification without disrupting gradient flow. Second, we model cross-view geometric discrepancy as a structured shift in transformation space and introduce View-Conditioned Low-Rank Adaptation (VC-LoRA). Inspired by parameter-efficient fine-tuning techniques [1, 7], we inject view-aware low-rank residual branches into feed-forward layers to modulate transformations conditioned on view context. This design enables specialization for aerial and ground inputs without duplicating backbone parameters. By jointly modeling activation sparsification and transformation modulation, the proposed framework addresses foreground clutter and view-dependent geometric shifts specific to AG-ReID.

Our contributions are summarized as follows:

- We introduce the IGS module, which employs a soft masking mechanism to explicitly attenuate background noise guided by identity semantics, ensuring robust feature interaction.
- We propose the VC-LoRA module, which dynamically modulates the weights of the feed-forward networks in the parameter space to alleviate transformation conflicts caused by extreme view discrepancies.
- We present VIGA, an end-to-end framework that jointly addresses geometric distortion and clutter, achieving state-of-the-art performance on challenging benchmarks such as AG-ReID.v2 [13] and LAGPeR [18].

2 Related Work

2.1 General Person Re-Identification

Building upon well-established ground surveillance infrastructure, advanced person re-identification (Re-ID) algorithms have made significant progress. Early methods predominantly relied on convolutional neural networks to extract robust features. Foundational works like ResNet [5] laid the groundwork, while subsequent approaches such as PCB [16] and MGN [17] introduced part-level slicing strategies to achieve fine-grained alignment. OSNet [22] further advanced this by aggregating multi-scale features to handle diverse resolutions. With the advent of vision transformers [3], attention-based models such as the pioneering TransReID [6] and the recently proposed DRFormer [15] have pushed the performance boundaries by capturing long-range dependencies and incorporating side information. However, with the proliferation of UAVs augmenting traditional surveillance systems, Re-ID research is increasingly extending to aerial-view scenarios. Initial efforts in this direction focused on aerial-aerial retrieval tasks using datasets such as PRAI-1581 [21]. While these works address resolution and scale variations, they primarily operate within a homogeneous top-down view setting. Consequently, they lack the capability to handle the extreme cross-view geometric discrepancies inherent in cross-platform matching, necessitating the development of specialized algorithms for heterogeneous aerial-ground networks.

2.2 Aerial-Ground Person Re-Identification

More recently, attention has shifted towards connecting aerial views with traditional ground systems, termed aerial-ground person ReID [12, 13]. In this setting, query and gallery images originate from fundamentally different viewpoints—typically high-altitude top-down views versus ground-level horizontal perspectives. This setup introduces severe spatial and semantic discrepancies, where identity-preserving cues such as face and clothing texture are often distorted or occluded by complex background clutter.

Several approaches have been proposed to mitigate these challenges. VDT [20] explicitly subtracts view-specific features from global embeddings to recover identity information. SeCap [18] leveraged prompt learning strategies to inject

view-related priors at the input level. Other works, such as LATex [8] and text-based retrieval methods [23], explore multi-modal knowledge to assist visual alignment. To address background clutter, DTST [19] proposes a dynamic token selection strategy to filter out irrelevant regions. Despite these advances, existing methods still face limitations in balancing structural integrity and noise suppression. Static backbone weights (e.g., in VDT, SeCap) struggle to fit conflicting geometric distributions. Furthermore, while hard selection methods like DTST effectively reduce computation, they inevitably disrupt the spatial topology of feature maps by physically removing tokens, hindering the preservation of geometric context. Concurrently, GeoReID [4] attempts to explicitly rectify geometry-induced distortions, yet its heavy reliance on auxiliary camera metadata restricts its deployment flexibility. Conversely, traditional soft attention mechanisms maintain structure but often fail to completely eliminate residual noise from high-response background areas.

To overcome these limitations, we propose a unified framework integrating VC-LoRA and IGS. VC-LoRA performs instance-adaptive geometric transformation in the parameter space. Complementarily, IGS employs an identity-guided soft masking mechanism. Unlike hard selection, IGS preserves the full spatial structure while explicitly attenuating background noise to a negligible level, ensuring robust and spatially consistent feature learning across aerial and ground domains.

3 Method

3.1 Overview

Given a training set $\mathcal{D} = \{(x_i, y_i, v_i)\}_{i=1}^{N}$, where x_i denotes the input image, y_i the identity label, and v_i the view label (aerial or ground), our goal is to learn view-invariant identity representations for cross-view retrieval. As illustrated in Fig. 1, our framework addresses two fundamental challenges in AG-ReID: severe foreground–background imbalance in aerial imagery and heterogeneous geometric transformations between aerial and ground views.

To tackle these challenges, we design a three-stage architecture. First, the backbone extracts global and local features while disentangling identity and view representations. Second, an IGS module suppresses background-dominated tokens to refine token interactions. Third, a View-Conditioned Decoder aggregates features and adaptively modulates the transformation according to the viewing geometry, where the proposed VC-LoRA further enhances the decoder by adaptively modeling view-dependent geometric shifts. These components jointly enable robust cross-view representation learning.

3.2 Feature Extraction and Disentanglement

We employ a Vision Transformer (ViT) encoder and adopt the VDT-style identity-view disentanglement strategy [20]. Given an input image x, the backbone outputs a sequence of features containing a global classification token f_{cls}, a specialized view token f_{view}, and a set of local patch tokens $F_{local} = \{f_1, f_2, \ldots, f_L\}$.

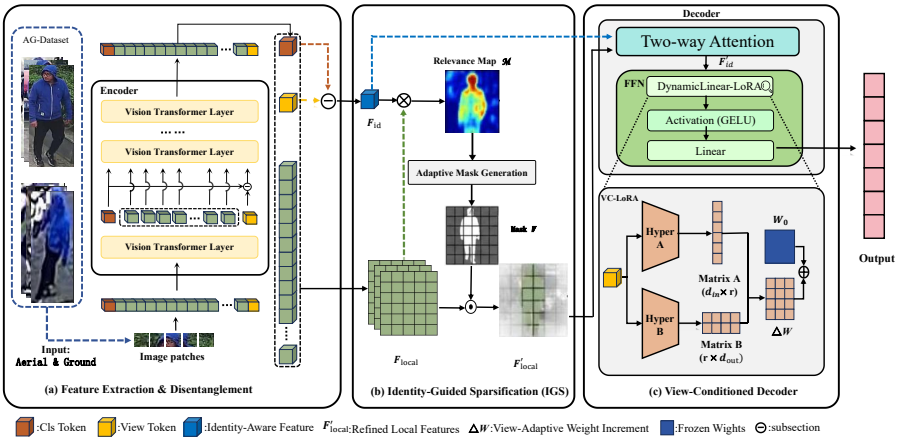


Fig. 1: Architecture of the proposed VIGA framework. (a) The ViT encoder with VDT-style identity-view disentanglement extracts the `cls` token, `view` token, and local features. (b) The IGS module utilizes the identity-aware feature (F_{id}), decoupled from the `cls` and `view` tokens, to compute relevance maps and actively suppresses background noise in the local features via adaptive masking. (c) VC-LoRA aggregates discriminative cues by interacting the identity-aware feature with the purified local features. Crucially, the FFN is dynamically modulated by the view token via VC-LoRA to adapt to view-dependent geometric shifts.

Identity-View Disentanglement. In AG-ReID, the global token f_{cls} inherently entangles identity information with viewpoint bias. To obtain a pure identity representation, we inherit the VDT-style subtraction-based disentanglement strategy [20] and use it as the identity-view separation basis for VIGA. Specifically, for the l-th transformer layer, we treat the view token $f_{view}^{(l)}$ as a prototype of the current viewpoint's characteristics and subtract it from the corresponding global representation:

$$F_{id}^{(l)} = \mathrm{Norm}(f_{cls}^{(l)} - f_{view}^{(l)}), \quad l = 1, \ldots, N_{blk}, \tag{1}$$

where $\mathrm{Norm}(\cdot)$ denotes LayerNorm and N_{blk} is the number of transformer blocks. Here the superscript (l) indexes the transformer layer. This operation forces each layer-wise identity feature $F_{id}^{(l)}$ to retain identity-consistent semantics while stripping away view-specific attributes. In the following sections, F_{id} denotes the identity feature produced by the final disentanglement stage. To ensure effective separation, we impose an orthogonality constraint on these decoupled features, which will be detailed in Sec. 3.5.

3.3 Identity-Guided Sparsification

Aerial images typically suffer from overwhelming background clutter, which can dominate the feature interaction in standard Transformers. To address this, we propose the IGS module. Unlike previous methods (e.g., DTS [19]) that perform hard token removal and thus disrupt the spatial topology of the feature map, IGS employs a soft masking mechanism. The key motivation is to use F_{id} as a top-down semantic guide: the mask is determined by identity relevance rather than by a heuristic threshold over token magnitude. Thus, IGS explicitly suppresses identity-irrelevant activations according to identity semantics.

Relevance Scoring. We first measure the semantic relevance between the decoupled identity feature F_{id} and each local patch $f_i \in F_{local}$:

$$s_i = \frac{F_{id}^{\top} f_i}{\|F_{id}\| \|f_i\|}, \quad \forall i \in \{1, \ldots, L\}, \tag{2}$$

This score vector $\mathbf{s}$ indicates the likelihood of a patch belonging to the pedestrian's body.

Soft Masking and Attenuation. Previous token pruning methods [10, 14, 19] typically discard low-importance tokens through hard removal, which breaks the spatial topology of the feature map and may hinder subsequent feature interactions. Hard masking keeps the token layout but creates abrupt feature cutoffs around foreground boundaries, which can remove fine-grained body-transition cues. To preserve the spatial structure while suppressing irrelevant activations, we adopt a soft masking strategy.

We first compute the semantic similarity between the identity feature and each local patch, and select the indices of the top-k most relevant patches, denoted as $\mathcal{I}_{\text{top}}$. Instead of removing the remaining tokens, we apply a deterministic attenuation mask $\mathcal{M}$ that selectively scales token activations while keeping the full token sequence intact:

$$\mathcal{M}_i = \begin{cases} 1.0 & \text{if } i \in \mathcal{I}_{\text{top}} \\ \epsilon & \text{otherwise} \end{cases}, \tag{3}$$

where ϵ is a strong attenuation factor (set to 10^{-3} in implementation). The refined local features are then computed element-wise for each patch:

$$\hat{f}_i = \mathcal{M}_i \cdot f_i, \quad \forall i \in \{1, 2, \ldots, L\}, \tag{4}$$

where L is the total sequence length of local patches. The positional encoding is not added back to the attenuated token values; instead, positional information is fed in parallel to the subsequent two-way attention decoder as spatial queries. This preserves spatial awareness without re-amplifying background activations.

3.4 View-Conditioned Decoder

To further integrate the identity-aware feature F_{id} with the refined local tokens $\hat{F}_{local}$, we employ a View-Conditioned Decoder that aggregates features through

a two-way attention mechanism. The decoder treats the identity representation as the query and the local tokens as key-value pairs, enabling bidirectional interaction between global identity cues and spatial features. The spatial positional queries are provided in parallel to the decoder, so the attenuation in IGS suppresses token activations without discarding their layout. This interaction allows the model to selectively focus on discriminative regions when constructing the final representation. However, the feed-forward networks (FFNs) within the decoder apply the same transformation to all samples. In AG-ReID, aerial and ground views exhibit drastically different geometric structures, such as top-down compression versus upright human silhouettes. Applying a single shared transformation forces the model to approximate heterogeneous geometric mappings, which limits the representational capacity of the decoder. To address this issue, we introduce a view-conditioned adaptation mechanism within the FFN layers.

View-Conditioned Low-Rank Adaptation. Specifically, we introduce the VC-LoRA module to dynamically adjust the transformation according to the viewing geometry. Let f_{view} denote the view token extracted by the backbone. Instead of using a fixed FFN weight matrix W_0, we augment it with a view-conditioned residual transformation

$$W(v) = W_0 + \alpha \Delta W, \tag{5}$$

where ΔW represents a view-dependent update and α is a learnable scaling factor.

To maintain computational efficiency, the residual update is parameterized using a low-rank decomposition

$$\Delta W = AB, \tag{6}$$

where $A \in \mathbb{R}^{d_{in} \times r}$ and $B \in \mathbb{R}^{r \times d_{out}}$ with $r \ll \min(d_{in}, d_{out})$. The low-rank factors are generated from the view token via lightweight hyper-networks

$$A = \Phi_A(f_{view}), \quad B = \Phi_B(f_{view}). \tag{7}$$

This design enables the decoder to capture view-specific transformation shifts while introducing only a small number of additional parameters. For stable optimization, the balancing coefficient α is initialized to 1.0 in our implementation. The low-rank residual branch follows the standard LoRA initialization practice, where the B branch is initialized to make the residual approximately zero at the beginning of training. Thus, the model starts close to the original pretrained transformation while retaining a nonzero coefficient for learning view-aware adaptations.

3.5 Optimization

The proposed framework is trained in an end-to-end manner. The overall objective function comprises three components: the identity loss $\mathcal{L}_{id}$, the view classification loss $\mathcal{L}_{view}$, and the orthogonality loss $\mathcal{L}_{orth}$.

Identity Loss. To ensure discriminative feature learning, we employ the standard identity loss $\mathcal{L}_{id}$, which combines the Cross-Entropy loss and the Triplet loss with hard mining [11] on the final aggregated embedding.

View Classification Loss. To explicitly force the decoupled view token f_{view} to capture viewpoint characteristics, we apply a view classification loss. It is formulated as a standard Cross-Entropy loss supervised by the ground-truth view labels:

$$\mathcal{L}_{view} = -\frac{1}{B}\sum_{i=1}^{B} \log p(v_i | f_{view}^{(i)}), \tag{8}$$

where B is the batch size, and $p(v_i | f_{view}^{(i)})$ denotes the predicted probability of the true view label $v_i \in \{\text{aerial}, \text{ground}\}$ for the i-th sample.

Orthogonality Loss. To ensure thorough decoupling between the identity and view-specific semantics, we apply an orthogonality constraint. Specifically, we minimize the squared cosine similarity between the two decoupled representations:

$$\mathcal{L}_{orth} = \left(\frac{\langle F_{id}, f_{view} \rangle}{\|F_{id}\|_2 \|f_{view}\|_2} \right)^2, \tag{9}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\| \cdot \|_2$ represents the L_2 norm. This formulation strictly forces the extracted features to be orthogonal in the angular space, effectively preventing the entanglement of view bias within the identity representation.

Overall Objective. The total loss function is formulated as the weighted sum of the above terms:

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \lambda_1 \mathcal{L}_{view} + \lambda_2 \mathcal{L}_{orth}, \tag{10}$$

where λ_1 and λ_2 are hyper-parameters balancing the contribution of the respective loss terms.

4 Experiments

4.1 Experimental Settings

Datasets. To strictly evaluate the effectiveness of VIGA in real-world aerial-ground scenarios, we conduct experiments on two large-scale AG-ReID datasets: LAGPeR [18] and AG-ReID.v2 [13].

- **LAGPeR :** This dataset contains 63,841 images of 4,231 identities collected from 21 cameras (7 aerial and 14 ground). Following the standard protocol, 2,708 identities are used for training, and the remaining 1,523 identities are reserved for testing. The evaluation is conducted under three challenging settings: aerial-to-ground ($A \rightarrow G$), ground-to-aerial ($G \rightarrow A$), and the comprehensive ground-to-aerial-plus-ground ($G \rightarrow A + G$).

– **AG-ReID.v2 :** This dataset consists of 100,502 images from 1,605 identities, captured by three distinct platforms: aerial drones (A), ground CCTV (C), and wearable devices (W). We use 807 identities for training and 798 for testing. The performance is evaluated under four cross-platform settings: $A \rightarrow C$, $C \rightarrow A$, $A \rightarrow W$, and $W \rightarrow A$, which pose challenges due to extreme view discrepancies.

Evaluation Metrics. We adopt the standard Rank-1 accuracy and mean Average Precision (mAP) to quantitatively evaluate the retrieval performance. Rank-1 measures the matching accuracy of the top-1 retrieval result, while mAP evaluates the overall retrieval quality across all correct matches.

Implementation Details. All experiments are implemented using PyTorch on NVIDIA A100 GPU.

– **Backbone & Input:** We employ the Vision Transformer (ViT-B/16) [3] pre-trained on ImageNet [2] as our feature extractor. All input images are resized to 256×128 pixels.
– **Training Strategy:** The model is trained for 120 epochs using the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. The batch size is set to 64, containing 16 identities with 4 images per identity. We utilize a cosine learning rate decay schedule, starting from an initial learning rate of 8×10^{-3}.
– **Hyper-parameters:** For our proposed modules, we set the activation preservation ratio ρ in the IGS module to 0.875 by default. In the VC-LoRA module, the rank r of the dynamic adaptation matrices is set to 16, and the balancing coefficient α is initialized to 1.0.

4.2 Comparison with State-of-the-Art

As shown in Table 1 and Table 2, we evaluate VIGA against state-of-the-art competitors on the LAGPeR and AG-ReID.v2 datasets. For a fair comparison, we focus exclusively on Transformer-based architectures. Specifically, we compare single-view ReID methods (ViT, CLIP-ReID, TransReID) and specialized AG-ReID methods (VDT, MIP, AG-ReID, and SeCap), all utilizing ViT as the backbone.

Results on LAGPeR. Table 1 reports the performance on the challenging LAGPeR dataset. Our method achieves competitive and often state-of-the-art performance across the evaluated protocols. Compared to the baseline ViT, our method improves Rank-1 accuracy by +4.8% and +4.11% on the $A \rightarrow G$ and $G \rightarrow A$ settings, respectively. Furthermore, compared to the recent method SeCap, our approach achieves 43.47% Rank-1 accuracy on the aerial-to-ground query task. Notably, in the most challenging $G \rightarrow A + G$ setting, which tests the model's comprehensive retrieval ability across mixed views, our method outperforms VDT, AG-ReID, and SeCap, demonstrating that our view-conditioned adaptation effectively mitigates geometric distortion inherent in aerial views.

Results on AG-ReID.v2. Table 2 presents the comparison on the AG-ReID.v2 dataset, covering interactions between Aerial (A), CCTV (C), and

Table 1: Performance comparison on the LAGPeR dataset. 'A → G' denotes that the Aerial view is the query, 'G → A' denotes that the Ground view is the query, and 'G → A + G' indicates that the gallery contains images from both the Aerial and Ground view. CLIP-ReID* indicates using OLP and SIE in Clip-ReID. MIP† represents the re-implementation for the AG-PReID. AG-ReID‡ indicates removing the attributes branch of the AG-ReID method. The best performance is in **bold**.

Method	**Backbone**	**LAGPeR**					
		$A \to G$		$G \to A$		$G \to A+G$	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ViT	ViT	38.67	27.25	32.04	30.69	18.88	15.31
TransReID	ViT	38.80	28.80	33.00	32.10	22.90	18.80
CLIP-ReID	CLIP	24.40	17.60	21.30	20.80	12.30	10.20
CLIP-ReID*	CLIP	23.10	17.50	20.00	20.30	9.00	8.40
MIP†	ViT	39.30	29.30	33.90	32.60	21.00	17.30
AG-ReID‡	ViT	40.48	28.89	32.96	31.91	22.03	17.89
VDT	ViT	40.15	28.97	33.55	31.98	19.50	16.45
SeCap	ViT	41.79	30.37	35.26	33.42	24.39	19.24
VIGA(Ours)	ViT	**43.47**	**31.52**	**36.05**	**34.44**	**25.11**	**20.18**

Table 2: Performance comparison on the AG-ReID.v2 dataset. C represents CCTV, W represents wearable devices, and A represents aerial views. The best results are highlighted in **bold**, while the second-best results are underlined.

Method	**Backbone**	$A \to C$		$C \to A$		$A \to W$		$W \to A$	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
BoT	ViT	85.40	77.03	84.65	75.90	89.77	80.48	84.65	75.90
AG-ReIDv1	ViT	87.70	79.00	87.35	78.24	**93.67**	83.14	87.73	79.08
VDT	ViT	86.46	79.13	86.14	78.12	90.00	82.21	85.26	78.52
AG-ReIDv2	ViT	<u>88.77</u>	80.72	87.86	78.51	<u>93.62</u>	**84.85**	**88.61**	80.11
SeCap	ViT	88.12	<u>80.84</u>	<u>88.24</u>	<u>79.99</u>	91.44	84.01	87.56	<u>80.15</u>
VIGA(Ours)	ViT	**88.88**	**82.37**	**88.96**	**81.51**	90.36	<u>84.39</u>	<u>87.95</u>	**81.75**

Wearable (W) views. Our method shows consistent improvements, particularly in the Aerial-CCTV scenarios ($A \leftrightarrow C$), which involve the most drastic viewpoint changes. Specifically, we achieve 88.88% Rank-1 in $A \to C$ and 88.96% in $C \to A$, surpassing both the specific AG-ReID v1/v2 methods and SeCap. Although our method does not utilize the additional attribute annotations used by AG-ReID, it still achieves competitive results on the Wearable splits ($A \leftrightarrow W$) and secures the best results on the core Aerial-Ground tasks. These results indicate that by adaptively recalibrating parameters to match the current view, our method learns more robust view-invariant features than static Transformer baselines, effectively reducing view bias in cross-platform retrieval.

Table 3: Ablation study of different components on LAGPeR and AG-ReID.v2 datasets. We report Rank-1 and mAP scores (%). The best results are highlighted in **bold**.

Setting	LAGPeR						AG-ReID.v2							
	$A \to G$		$G \to A$		$G \to A+G$		$A \to C$		$C \to A$		$A \to W$		$W \to A$	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Baseline	41.79	30.37	35.26	33.42	24.39	19.24	88.12	80.84	88.24	79.99	**91.44**	84.01	87.56	80.15
+ VC-LoRA	42.78	31.00	35.36	34.00	24.39	19.90	88.12	82.01	88.18	81.20	90.86	84.17	**88.29**	81.35
+ **VC-LoRA + IGS**	**43.47**	**31.52**	**36.05**	**34.44**	**25.11**	**20.18**	**88.88**	**82.37**	**88.96**	**81.51**	90.36	**84.39**	87.95	**81.75**

4.3 Ablation Study

To assess the individual contributions of each component in VIGA, we conduct a comprehensive ablation study on the LAGPeR and AG-ReID.v2 datasets. As shown in Table 3, we start with a ViT-based baseline and progressively integrate the **VC-LoRA** and **IGS** modules.

Compared to the baseline, integrating VC-LoRA brings consistent performance improvements across most settings. Specifically, on the LAGPeR $A \to G$ protocol, VC-LoRA improves Rank-1 from 41.79% to 42.78% and mAP from 30.37% to 31.00%. Similarly, on the AG-ReID.v2 $A \to C$ setting, mAP increases from 80.84% to 82.01%. These results confirm the effectiveness of dynamic parameter adaptation in compensating for cross-view geometric deformations. Interestingly, the improvement in mAP is often more pronounced than Rank-1 in complex scenarios, indicating that VC-LoRA helps in learning a more consistent feature space.

Adding the IGS module on top of VC-LoRA yields further improvements, ensuring the model focuses on discriminative regions. The full model achieves the strongest overall performance across the evaluated scenarios. Notably, in the LAGPeR $G \to A+G$ setting, IGS pushes the Rank-1 accuracy to 25.11% and mAP to 20.18%. On AG-ReID.v2, the combination achieves a peak Rank-1 of 88.88% for $A \to C$. This demonstrates that IGS effectively suppresses background clutter in aerial imagery by leveraging identity-guided masking, thus enhancing cross-view alignment.

Overall, the ablation study verifies the complementary effects of VC-LoRA and IGS. While VC-LoRA resolves geometric distortions in the weight space, IGS provides spatial-level refinement by softly masking irrelevant background noise. Their combination allows VIGA to maintain robust performance across diverse matching protocols, especially under severe viewpoint and occlusion conditions.

To further justify the masking design in IGS, we compare the proposed soft-masking strategy with a hard-masking alternative. Token pruning or hard masking removes low-relevance responses more aggressively, but it can break the 2D token grid or introduce abrupt feature cutoffs around human boundaries. In contrast, soft masking attenuates identity-irrelevant activations while preserving all token positions, which maintains spatial continuity for downstream geometric alignment. As shown in Table 4, soft masking is not uniformly superior on every

Table 4: Comparison between hard masking and the proposed soft masking. We report Rank-1 and mAP scores (%).

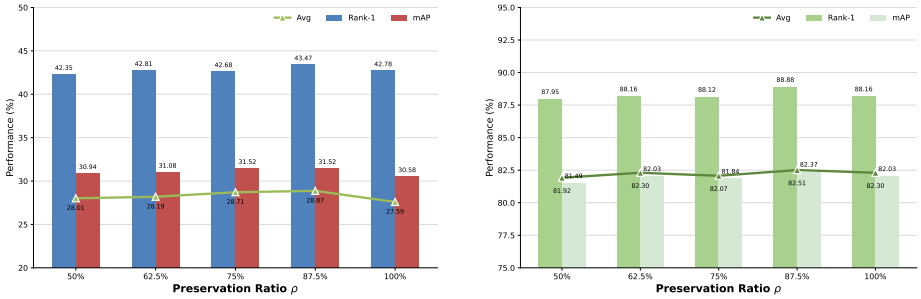
Masking Strategy	LAGPeR						AG-ReID.v2							
	$A \to G$		$G \to A$		$G \to A+G$		$A \to C$		$C \to A$		$A \to W$		$W \to A$	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Hard masking	41.76	30.65	34.60	33.67	**25.25**	19.89	88.67	**82.40**	87.85	81.36	**90.77**	84.10	87.63	81.07
Soft masking (VIGA)	**43.47**	**31.52**	**36.05**	**34.44**	25.11	**20.18**	**88.88**	82.37	**88.96**	**81.51**	90.36	**84.39**	**87.95**	**81.75**

individual metric, but it provides stronger overall performance across the two benchmarks. On AG-ReID.v2, soft masking improves $C \to A$ and $W \to A$ while remaining competitive on $A \to C$ and $A \to W$. On LAGPeR, soft masking improves $A \to G$ and $G \to A$, and it also improves mAP on $G \to A + G$ while remaining competitive in Rank-1.

4.4 Hyperparameter Sensitivity

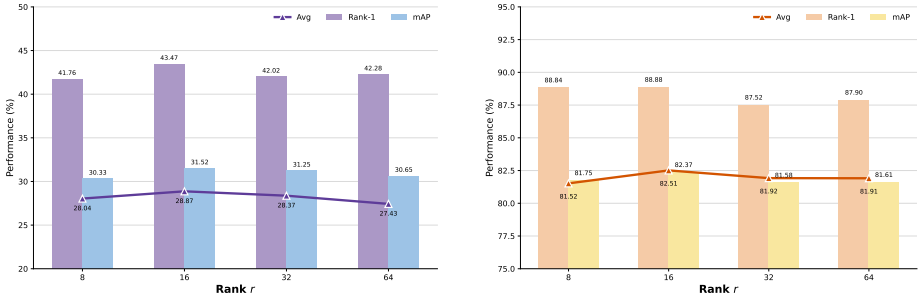
Impact of the Preservation Ratio ρ. In the IGS module, the hyperparameter ρ determines the proportion of tokens that retain their full activation for spatial alignment, while the rest are softly suppressed. To study its effect, we vary ρ within the range of $\{50\%, 62.5\%, 75\%, 87.5\%, 100\%\}$. As shown in Figure 2, the performance trend on both datasets is consistent: it ascends as ρ increases and drops slightly when retaining all tokens. Notably, both LAGPeR and AG-ReID.v2 achieve optimal performance at $\rho = 87.5\%$. A lower ratio (e.g., 50%) leads to the loss of discriminative cues, degrading the feature representation. Conversely, retaining all tokens ($\rho = 100\%$) introduces irrelevant background noise and occlusions, which hinders effective cross-view matching. These results demonstrate that softly masking a small portion (approx. 12.5%) of the least informative patches effectively suppresses noise while preserving semantic integrity.

Impact of the Rank r. We further investigate the influence of the rank dimension r in the VC-LoRA module, which controls the number of trainable parameters for geometric adaptation. We evaluate the model performance by varying r within the set $\{8, 16, 32, 64\}$. As illustrated in Figure 3, the performance on both LAGPeR and AG-ReID.v2 datasets exhibits a consistent trend: it initially improves as r increases and then degrades. Specifically, the model achieves peak performance at $r = 16$. A smaller rank (e.g., $r = 8$) lacks sufficient capacity to model complex view-specific geometric variations, resulting in underfitting. Conversely, increasing the rank to 32 or 64 leads to a performance drop, likely due to overfitting caused by the introduction of excessive parameters relative to the dataset scale. Therefore, we adopt $r = 16$ as the optimal setting to balance adaptation flexibility and parameter efficiency.



(a) LAGPeR **(b)** AG-ReID.v2

Fig. 2: Sensitivity analysis of the Preservation Ratio ρ on different datasets. The results consistently show that preserving a high ratio of tokens (e.g., 87.5%) yields optimal performance by balancing noise suppression and information integrity.



(a) LAGPeR **(b)** AG-ReID.v2

Fig. 3: Sensitivity analysis of the Rank r on different datasets. Both datasets achieve peak performance at $r = 16$, demonstrating that a moderate rank is sufficient to capture view-specific geometric variations without overfitting.

5 Conclusion

We propose VIGA, a novel framework for aerial-ground person re-identification by jointly addressing foreground-background imbalance and AG-ReID-specific aerial-ground geometric discrepancies. Specifically, we introduce identity-guided sparsification to suppress irrelevant token activations while preserving spatial structure, and view-conditioned low-rank adaptation to modulate backbone transformations under different viewpoints. Extensive experiments on challenging AG-ReID benchmarks demonstrate consistent improvements over strong transformer-based baselines. These results highlight the importance of combining activation-level refinement with view-aware parameter adaptation for robust cross-view representation learning.

Acknowledgment

This work is supported in part by the National Key R&D Program of China (No. 2025YFF0514700), in part by the Chongqing Research Institute Performance Incentive Guidance Special Project (No. CSTB2025JXJL-YFX0044), and in part by the Chongqing Municipal Education Commission (No. KJZD-M202400603, No. KJQN202500612).

References

1. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
4. Hambarde, K.A., Proença, H.: Rectifying geometry-induced similarity distortions for real-world aerial-ground person re-identification. arXiv preprint arXiv:2601.21405 (2026)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 770–778 (2016)
6. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15013–15022 (2021)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
8. Hu, X., Wang, Y., Zhang, P.: Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. arXiv preprint arXiv:2503.23722 (2025)
9. Li, S., Sun, L., Li, Q.: Clip-reid: Exploiting vision-language model for image re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
10. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800 (2022)
11. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: 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 (CVPRW). pp. 0–0 (2019)
12. Nguyen, H., Nguyen, K., Sridharan, S., Fookes, C.: Aerial-ground person re-id. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11165–11175 (2023)
13. Nguyen, H., Nguyen, K., Sridharan, S., Fookes, C.: Ag-reid. v2: Bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security (T-IFS) (2024)

14. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 13937–13949 (2021)
15. Shu, Y., Zhan, P., Yang, H.: Drformer: A dual-regularized bidirectional transformer for person re-identification. arXiv preprint arXiv:2602.01059 (2026)
16. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (pcb). In: European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
17. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia (MM). pp. 274–282 (2018)
18. Wang, S., Wang, Y., Wu, R., Jiao, B., Wang, W., Wang, P.: Secap: self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22119–22128 (2025)
19. Wang, Y., Pishgar, M.: Dynamic token selective transformer for aerial-ground person re-identification. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
20. Zhang, Q., Wang, L., Patel, V.M., Xie, X., Lai, J.: 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 (CVPR). pp. 1234–1244 (2024)
21. Zhang, S., Wei, Y., Zhang, Q., Zhang, Y., Guo, Y.: Person re-identification in aerial imagery. IEEE Transactions on Multimedia (T-MM) **22**(8), 2132–2143 (2019)
22. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3702–3712 (2019)
23. Zhou, X., Wu, Y., Ma, J., Wang, W., Cao, M., Ye, M.: Text-based aerial-ground person retrieval. arXiv preprint arXiv:2511.08369 (2025)