QVAM: Query-guided View-aware Adaptive Modulation for Aerial-Ground Person Re-Identification

Mengfei Zhou and Lin Wan*

School of Computer Science, China University of Geosciences, Wuhan, China
{1202511217,wanlin}@cug.edu.cn

Abstract. Aerial–Ground Person Re-Identification (AGPReID) matches identities across unmanned aerial vehicles (UAVs) and ground cameras but suffers from extreme viewpoint gaps. Most existing methods rely on coarse binary aerial/ground labels and rigid orthogonality to disentangle view and identity features. We identify an important limitation of prior methods: the oversimplified binary-label design cannot fully capture continuous aerial viewpoint shifts. Moreover, rigid orthogonality constraints may further suppress identity cues. To address these issues, we propose Query-guided View-aware Adaptive Modulation (QVAM) for identity-preserving cross-view matching. Specifically, a View-aware Decoder (VAD) distills fine-grained viewpoint cues from patch tokens using learnable view queries. Guided by these cues, Adaptive Feature Modulation (AFM) predicts query-conditioned masks to suppress view-biased responses while preserving identity discrimination. A Cross-View Prototype Alignment (CVPA) loss further aligns modulated features at batch and memory levels with dual-view memory banks. Extensive experiments on AG-ReID, AG-ReIDv2, and CARGO show that QVAM achieves state-of-the-art performance, improving the previous best results by +10.26% Rank-1/+10.83% mAP on CARGO-ALL and +2.62% Rank-1/+2.37% mAP on AG-ReID A→G. The code is available at `https://github.com/Sakuraandroxy/QVAM`.

Keywords: Aerial-Ground Person Re-Identification · UAV Surveillance · Continuous Viewpoint Modeling · Adaptive Feature Modulation · Prototype Alignment

1 Introduction

Person Re-Identification (ReID) aims to retrieve images of a target person from a large gallery given a query image and plays a pivotal role in intelligent surveillance, security, and suspect tracking [42, 49]. Driven by deep learning [9], recent studies have achieved strong performance on standard benchmarks by addressing challenges such as occlusion [24, 48, 55], domain generalization [3, 14, 31],

* Corresponding author.

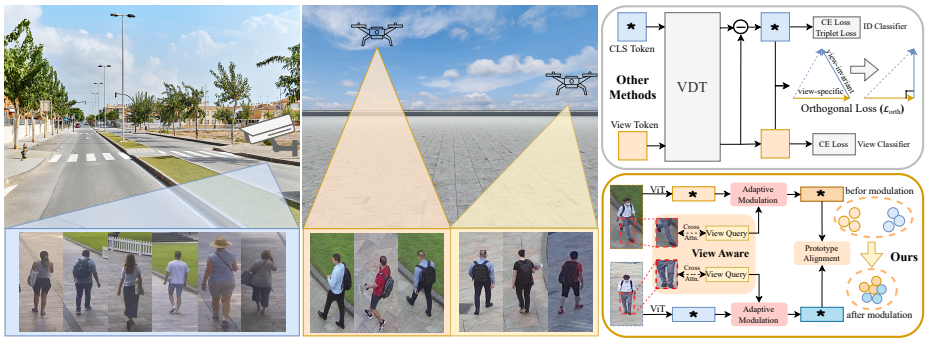


Fig. 1: Challenges in AGPReID and comparison to existing methods. (Left) Ground views are relatively consistent. (Middle) Aerial views exhibit substantial *intra-view diversity* due to altitude and depression-angle variations. (Right) Comparison to existing methods: (Top) Disentanglement-based methods (e.g., VDT [46]) enforce orthogonality ($\mathcal{L}_{\text{orth}}$) between view-specific and view-invariant features. (Bottom) Our QVAM adaptively modulates `[CLS]` token with viewpoint cues, which effectively aligns these features.

cross-modality matching [2, 18], and cloth-changing scenarios [7, 20, 29]. However, ground-camera surveillance is inherently limited by fixed deployments, restricted fields of view, and frequent occlusions. To extend coverage, UAVs have recently been integrated into surveillance systems, enabling aerial–ground person retrieval (AGPReID). Despite its practical value, AGPReID remains challenging due to the drastic viewpoint gap between overhead UAV views and horizontal ground views, which induces severe feature misalignment (low intra-identity similarity and high inter-identity similarity). As a result, ReID models trained in conventional single-domain settings often generalize poorly in this heterogeneous scenario.

To mitigate the aerial–ground viewpoint discrepancy, existing approaches [36, 46] typically learn view-specific features under coarse binary aerial/ground supervision and enforce disentanglement with orthogonality constraints. However, as shown in Fig. 1, UAV altitude and depression-angle changes introduce *continuous* and fine-grained aerial viewpoint variations that binary view labels fail to capture, causing intra-view feature fragmentation and the resulting aerial–ground misalignment. Moreover, rigid orthogonality constraints can be overly restrictive: view-specific representations may still encode identity-discriminative cues, and forcing a hard partition may suppress shared semantics, degrading identity matching (see Sec. 4.5).

To address these challenges, we propose Query-guided View-aware Adaptive Modulation (QVAM), which uses learnable view queries to model continuous aerial viewpoint variations and performs identity-preserving feature modulation for cross-view matching. Specifically, unlike approaches relying on explicit view disentanglement, the View-aware Decoder (VAD) leverages learnable view queries to distill fine-grained viewpoint cues from local patch tokens, go-

ing beyond coarse binary supervision. Guided by these cues, the Adaptive Feature Modulation (AFM) module predicts query-conditioned modulation masks to suppress viewpoint-biased responses while preserving identity-discriminative cues, avoiding rigid orthogonal decomposition. Finally, the Cross-View Prototype Alignment (CVPA) loss aligns the modulated features at both batch and memory levels via momentum-updated dual-view memory banks, further stabilizing aerial–ground alignment.

In summary, our contributions are as follows:

- We introduce QVAM, a query-guided framework that models continuous aerial viewpoint variations and enables identity-preserving cross-view matching.
- We propose VAD and AFM, which distill fine-grained viewpoint cues and perform query-conditioned feature modulation, respectively, alleviating the limitations of coarse binary supervision and rigid orthogonality constraints.
- We develop the CVPA loss that maintains dynamic view-specific prototypes for each identity and enforces alignment at both batch and memory levels, guiding distribution-level alignment of the modulated features.
- We achieve state-of-the-art performance on standard AGPReID benchmarks, improving upon the strongest prior results.

2 Related Work

2.1 View-Homogeneous ReID

View-homogeneous ReID refers to retrieving specific individuals from ground-based or aerial-based camera networks. Ground-based ReID has received extensive attention in both academia and industry. Several large-scale benchmarks, such as Market-1501 [49] and MSMT17 [39], have been established to facilitate research. Early approaches primarily relied on hand-crafted features [21, 41]. With the rapid development of deep learning, CNN-based [10, 17, 23, 32, 34, 42] and Transformer-based [11, 16, 30, 36, 37, 45, 46, 54] methods have become the mainstream. Although less explored, aerial-based ReID has also seen progress with valuable methods [15, 19, 35, 47] and datasets such as PRAI-1581 [47] and UAVHuman [19], focusing on matching within drone networks.

However, view-homogeneous networks still involve viewpoint issues, such as distinct feature discrepancies among frontal, dorsal, and lateral views in ground scenarios. Therefore, while some methods [1,40,43,44] have investigated resolving these homogeneous viewpoint differences, they are inapplicable to the drastic viewpoint disparities in heterogeneous aerial-ground camera networks.

2.2 View-Heterogeneous ReID

In this paper, View-heterogeneous ReID specifically refers to retrieving persons of interest across heterogeneous aerial-ground camera networks. The primary challenge lies in drastic viewpoint discrepancies that cause significant variations

in person appearance. Methods such as AG-ReID [25] utilize attribute labels to guide the model in learning view-invariant attribute features. Building upon this, AG-ReIDv2 [26] introduces an Elevated-view Attention Stream to leverage head features for assisting matching, yet neither effectively alleviates viewpoint discrepancies nor addresses the labor-intensive nature of attribute annotation. VIF [16] identifies the limitations of traditional image-level data augmentation (e.g., global flipping) in handling viewpoint discrepancies and proposes Patch-Level RotateMix, which probabilistically rotates and blends local image patches to facilitate view-invariant feature learning. VDT [46] utilizes view tokens to perform explicit view disentanglement on `[CLS]` token, optimized via orthogonal loss, while SeCap [36] extends this to learn view-invariant local features. However, they fail to account for diversity of viewpoints in AGPReID, as coarse binary view labels cannot capture fine-grained view-specific features. Furthermore, general feature modulation or disentanglement methods [12,50] and visual prompt tuning techniques [13,52] offer potential solutions. However, directly applying static prompts fails to address the diverse viewpoint discrepancies in AGPReID. Recent long-range biometric systems and aerial–ground ReID challenge reports, including FarSight [22], AG-ReID [27], and AG-VPReID [28], further highlight the practical importance of AGPReID.

3 Methodology

3.1 Overview

As illustrated in Fig. 2(a), we build QVAM on a Vision Transformer (ViT) backbone. Given an aerial or ground image, we tokenize it into non-overlapping patch embeddings and encode them with a ViT encoder, yielding a `[CLS]` token and patch tokens. The View-aware Decoder (VAD) uses learnable view queries to attend to the patch tokens and extract fine-grained viewpoint cues, producing view-aware query embeddings. Conditioned on these embeddings, the Adaptive Feature Modulation (AFM) module predicts query-conditioned modulation masks to adapt the `[CLS]` representation into a view-invariant embedding, enabling robust cross-view retrieval. The Cross-View Prototype Alignment (CVPA) loss further aligns modulated features at batch and memory levels with dual-view memory banks, effectively guiding the VAD and AFM modules.

3.2 View-aware Decoder

To distill fine-grained viewpoint cues, we introduce the **View-aware Decoder** (VAD). As illustrated in Fig. 2(b), this module adopts a standard Transformer decoder [33] architecture equipped with a set of learnable queries.

Formally, let $\mathcal{T}_{\text{patch}} \in \mathbb{R}^{B\times N\times C}$ denote the patch tokens extracted from the ViT backbone, where B, N, and C represent the batch size, the number of patches (e.g., 128), and the channel dimension, respectively. We initialize a set of learnable view queries, denoted as $\mathbf{Q} \in \mathbb{R}^{B\times L\times C}$, where L is a hyperparameter indicating the number of queries. The decoding process within the VAD

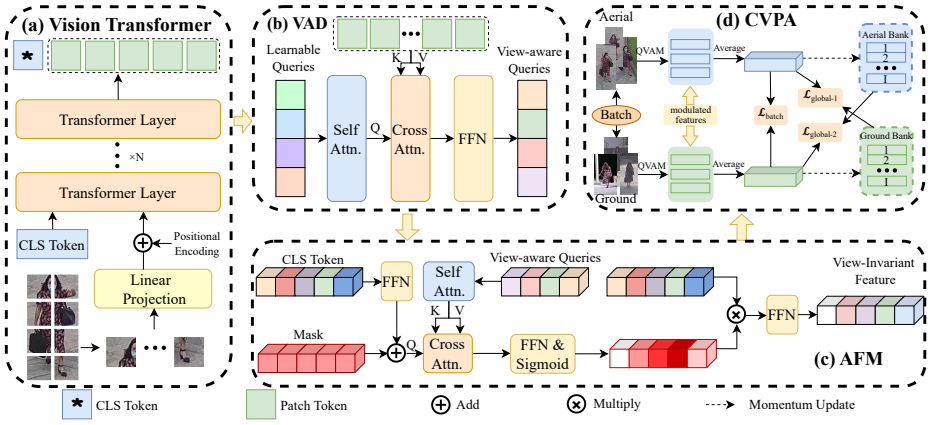


Fig. 2: QVAM architecture. (a) Vision Transformer (ViT) extracts a `[CLS]` token and patch-level tokens. (b) View-aware Decoder (VAD) distills fine-grained viewpoint cues with learnable view queries. (c) Adaptive Feature Modulation (AFM) predicts query-conditioned masks for `[CLS]` token modulation. (d) Cross-view Prototype Alignment (CVPA) loss aligns the modulated features across aerial and ground views using dual-view memory banks.

involves three sequential stages. First, the queries undergo a Multi-Head Self-Attention mechanism to model inter-query dependencies and initialize their semantic context. Subsequently, the queries interact with $\mathcal{T}_{\text{patch}}$ via Multi-Head Cross-Attention. Here, the view queries act as the queries (Q), while $\mathcal{T}_{\text{patch}}$ serves as the keys (K) and values (V), enabling the queries to aggregate fine-grained viewpoint cues from $\mathcal{T}_{\text{patch}}$. Finally, a Feed-Forward Network (FFN) projects the queries into a latent space suitable for the subsequent view-adaptive mask generation.

For a single VAD layer, the generation of the final view-aware queries, denoted as $\mathbf{Q}_{\text{view}}$, can be formulated as:

$$\mathbf{Q}_{\text{view}} = \text{FFN}\left(\text{MCA}\left(\text{MSA}(\mathbf{Q}), \mathcal{T}_{\text{patch}}, \mathcal{T}_{\text{patch}}\right)\right), \tag{1}$$

where $\text{MSA}(\cdot)$ denotes Multi-Head Self-Attention, and $\text{MCA}(Q, K, V)$ denotes Multi-Head Cross-Attention. Note that residual connections and layer normalization are applied following standard Transformer design (omitted in Eq. 1 for brevity).

To ensure that the queries capture comprehensive and non-redundant viewpoint cues, we introduce a Query Diversity Loss ($\mathcal{L}_{\text{div}}$). This mechanism is designed to prevent mode collapse by encouraging different queries to attend to distinct spatial regions (patches) of the input image, thereby maximizing the diversity of the extracted viewpoint cues.

Formally, let $\mathbf{A} \in \mathbb{R}^{B \times L \times N}$ denote the attention matrix output by the MCA layer in the VAD module, where B is the batch size, L is the number of queries, and N is the number of patch tokens. For the i-th image in a batch, let $\mathbf{a}_{i,j} \in \mathbb{R}^{N}$

represent the softmax-normalized attention vector of the j-th query towards all patch tokens. We impose an orthogonality constraint by minimizing the squared dot product between the attention vectors of any pair of distinct queries. The loss is formulated as:

$$\mathcal{L}_{\text{div}} = \frac{1}{BL(L-1)} \sum_{i=1}^{B} \sum_{j=1}^{L} \sum_{\substack{k=1 \\ k \neq j}}^{L} (\mathbf{a}_{i,j} \cdot \mathbf{a}_{i,k})^2 \ , \tag{2}$$

where $(\cdot)$ denotes the dot product operation. The resulting $\mathbf{Q}_{\text{view}}$ encapsulates rich viewpoint cues and is subsequently fed into the AFM module to guide the adaptive modulation of the `[CLS]` token.

3.3 Adaptive Feature Modulation

To adaptively suppress viewpoint biases and learn identity-preserving view-invariant features, we propose the **Adaptive Feature Modulation** (AFM) module which is illustrated in Fig. 2(c). Formally, let $\mathbf{x}_{\text{cls}} \in \mathbb{R}^{B \times C}$ denote the image-level `[CLS]` token extracted from the ViT backbone. We initialize a shared learnable mask token $\mathbf{m} \in \mathbb{R}^{1 \times C}$, which is broadcasted along the batch dimension to $\mathbb{R}^{B \times C}$ to form the initial mask token for each sample in the batch.

The mask generation process consists of two key phases: Identity Context Injection and View-Adaptive Refinement. First, to establish a holistic perception of the pedestrian identity before fine-grained modulation, we inject holistic identity context into the mask by fusing it with the `[CLS]` token. This initialization ensures that the mask prioritizes identity-level representations, providing a robust foundation for subsequent view-adaptive refinement. In this way, we obtain a context-aware intermediate mask $\mathbf{m}'$:

$$\mathbf{m}' = \mathbf{m} + \text{FFN}(\mathbf{x}_{\text{cls}}) \, . \tag{3}$$

Subsequently, the mask attends to the view-aware queries ($\mathbf{Q}_{\text{view}}$) to capture fine-grained local viewpoint cues, facilitating further refinement. Specifically, $\mathbf{Q}_{\text{view}}$ first undergoes Multi-Head Self-Attention (MSA) to model internal dependencies. Then, a Multi-Head Cross-Attention (MCA) mechanism is employed to refine the mask, where $\mathbf{m}'$ serves as the query (Q), and the view-aware queries serve as the keys (K) and values (V). Finally, the output is projected via a FFN and activated by a Sigmoid function to produce final view-adaptive mask $\mathbf{M} \in \mathbb{R}^{B \times C}$. This process is formulated as:

$$\mathbf{M} = \sigma\left(\text{FFN}\left(\text{MCA}\left(\mathbf{m}', \text{MSA}(\mathbf{Q}_{\text{view}}), \text{MSA}(\mathbf{Q}_{\text{view}})\right)\right)\right) \ , \tag{4}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation, which constrains mask values to the range $(0, 1)$ to serve as channel-wise scaling factors. Notably, in the absence of the VAD module, the view-adaptive refinement phase is bypassed. In this scenario, the final mask is directly derived from the intermediate mask solely based on image-level identity context, formulated as $\mathbf{M} = \sigma(\text{FFN}(\mathbf{m}'))$.

With the view-adaptive mask, we modulate the [CLS] token into *candidate* view-invariant features ($\mathbf{f}_{\text{inv}}$) and suppressed residuals ($\mathbf{f}_{\text{res}}$) via element-wise multiplication:

$$\mathbf{f}_{\text{inv}} = \text{FFN}(\mathbf{x}_{\text{cls}} \odot \mathbf{M}), \quad \mathbf{f}_{\text{res}} = \mathbf{x}_{\text{cls}} \odot (1 - \mathbf{M}), \tag{5}$$

where $\odot$ denotes the element-wise product. It is worth noting that while the modulation operation provides the *capacity* for feature selection, the actual view-invariance of $\mathbf{f}_{\text{inv}}$ is further encouraged by the subsequent CVPA loss (detailed in Sec. 3.4).

Mask Entropy Regularization. To avoid semantic ambiguity where the network fails to distinguish between channels requiring different modulation intensities (i.e., mask values clustering around 0.5), we introduce a Mask Entropy Regularization Loss ($\mathcal{L}_{\text{reg}}$). Unlike rigid binarization constraints, this objective serves as a soft regularization term to penalize high-entropy states. By pushing the mask values away from the uncertain equilibrium (0.5), it encourages the network to explicitly identify the semantic role of each channel, while still preserving the flexibility of fine-grained modulation. The loss is defined as:

$$\mathcal{L}_{\text{reg}} = -\frac{1}{B \times C} \sum_{i=1}^{B} \sum_{j=1}^{C} \left[\mathbf{M}_{i,j} \log(\mathbf{M}_{i,j} + \epsilon) + (1 - \mathbf{M}_{i,j}) \log(1 - \mathbf{M}_{i,j} + \epsilon) \right], \tag{6}$$

where $\epsilon = 10^{-7}$ is a small constant for numerical stability.

3.4 Cross-View Prototype Alignment Loss

To encourage view-invariant modulated features and bridge the distributional gap between aerial and ground modalities, as illustrated in Fig. 2 (d), we propose the **Cross-View Prototype Alignment** (CVPA) loss. Let the training set be denoted as $\mathcal{D} = \{(\mathbf{x}_k, y_k, v_k)\}_{k=1}^{N}$, where $\mathbf{x}_k$ is the k-th image, $y_k \in \{1, \ldots, I\}$ is its identity label, and $v_k \in \{\mathcal{A}, \mathcal{G}\}$ is its view label (Aerial or Ground). We construct Dual-View Memory Banks [5, 51], $\mathcal{M}^{\mathcal{A}}, \mathcal{M}^{\mathcal{G}} \in \mathbb{R}^{I \times C}$, to store view-specific identity prototypes in each view.

Prototype Update via Momentum. In each training iteration, we first compute the view-specific centroids for each identity i within the current batch. Let $\mathcal{B}_i^{\mathcal{A}}$ and $\mathcal{B}_i^{\mathcal{G}}$ denote the sets of feature vectors for identity i captured under the aerial and ground views within the current batch, respectively. The corresponding aerial centroid $\boldsymbol{\mu}_i^{\mathcal{A}}$ and ground centroid $\boldsymbol{\mu}_i^{\mathcal{G}}$ are calculated as:

$$\boldsymbol{\mu}_i^{\mathcal{A}} = \frac{1}{|\mathcal{B}_i^{\mathcal{A}}|} \sum_{\mathbf{f} \in \mathcal{B}_i^{\mathcal{A}}} \mathbf{f}, \quad \boldsymbol{\mu}_i^{\mathcal{G}} = \frac{1}{|\mathcal{B}_i^{\mathcal{G}}|} \sum_{\mathbf{f} \in \mathcal{B}_i^{\mathcal{G}}} \mathbf{f}, \tag{7}$$

where $\mathbf{f}$ refers to the modulated feature (before FFN) extracted by the AFM module.

Standard memory-based approaches typically update memory *after* loss computation to prevent information leakage. However, our $\mathcal{L}_{\text{mem}}$ aligns features with

prototypes from the *opposing* view, inherently circumventing self-matching risks. Leveraging this property, we update memory banks *prior* to loss calculation. This strategy effectively alleviates the cold-start problem, providing immediate supervision (see *supplementary material* for details). The momentum update [8] is formulated as:

$$\mathbf{p}_i^v \leftarrow \alpha \mathbf{p}_i^v + (1-\alpha)\boldsymbol{\mu}_i^v\,, \quad v \in \{\mathcal{A}, \mathcal{G}\}\,, \tag{8}$$

where $\mathbf{p}_i^v$ denotes the prototype of the i-th identity in $\mathcal{M}^v$ and α is a momentum hyperparameter.

Alignment Objectives. To mitigate the instability caused by limited samples in a single batch, we impose alignment constraints at both batch and memory levels. The batch-level alignment ($\mathcal{L}_{\text{batch}}$) minimizes the Euclidean distance between the aerial and ground centroids of the same identity within the current batch:

$$\mathcal{L}_{\text{batch}} = \sum_{i \in \mathcal{I}_{\text{batch}}} ||\boldsymbol{\mu}_i^{\mathcal{A}} - \boldsymbol{\mu}_i^{\mathcal{G}}||_2\,. \tag{9}$$

The memory-level alignment ($\mathcal{L}_{\text{mem}}$) aligns the batch centroids with the view-specific prototypes of the *opposing* view stored in the memory banks ($\mathbf{p}_i^{\mathcal{G}} \in \mathcal{M}^{\mathcal{G}}$ and $\mathbf{p}_i^{\mathcal{A}} \in \mathcal{M}^{\mathcal{A}}$):

$$\mathcal{L}_{\text{mem}} = \sum_{i \in \mathcal{I}_{\text{batch}}} \left(||\boldsymbol{\mu}_i^{\mathcal{A}} - \mathbf{p}_i^{\mathcal{G}}||_2 + ||\boldsymbol{\mu}_i^{\mathcal{G}} - \mathbf{p}_i^{\mathcal{A}}||_2\right)\,. \tag{10}$$

The total CVPA loss is defined as $\mathcal{L}_{\text{CVPA}} = \mathcal{L}_{\text{batch}} + \mathcal{L}_{\text{mem}}$.

3.5 Optimization Objectives

Initial `[CLS]` Supervision ($\mathcal{L}_{\text{cls}}$). To ensure discriminative feature learning in the backbone, we apply both identity-based Cross-Entropy and Triplet losses to the initial `[CLS]` token $\mathbf{x}_{\text{cls}}$.

View-Invariant Supervision ($\mathcal{L}_{\text{inv}}$). To prevent feature collapse during alignment and preserve discriminative identity information, we apply the Triplet loss to the intermediate modulated features (before FFN), while imposing both the identity-based Cross-Entropy loss and the Triplet loss on the final view-invariant features $\mathbf{f}_{\text{inv}}$.

The total optimization objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{inv}} + \gamma\mathcal{L}_{\text{CVPA}} + \beta\mathcal{L}_{\text{div}} + \xi\mathcal{L}_{\text{reg}}\,, \tag{11}$$

where β, γ, and ξ are hyperparameters that balance the query diversity loss, CVPA loss, and mask entropy regularization loss, respectively.

4 Experiment

4.1 Experimental Settings

Datasets. We evaluate our proposed method on three representative AGPReID benchmarks: AG-ReID [25], AG-ReIDv2 [26] and CARGO [46]. AG-ReID is the

first real-world dataset tailored for AGPReID, containing 21,398 images of 388 identities with 15 attribute annotations. Following the standard protocol, we report results under two cross-view settings, A→G and G→A, where A and G denote aerial and ground cameras, respectively. AG-ReIDv2 includes 100,502 images of 1,615 identities and provides four evaluation protocols: A→C, A→W, W→A, and C→A, where C and W denote CCTV and wearable cameras, respectively. CARGO is a large-scale synthetic benchmark with 108,563 images of 5,000 identities collected from five aerial and eight ground cameras. We evaluate under four protocols: ALL (mixed-view evaluation on the full gallery), A↔G, A↔A, and G↔G.

Evaluation Metrics. To comprehensively assess the performance of QVAM, we adopt standard evaluation metrics for person re-identification, including the Cumulative Matching Characteristic (CMC) at Rank-1 [6], mean Average Precision (mAP) [49], and mean Inverse Negative Penalty (mINP) [42]. Rank-1 measures top-1 retrieval accuracy, mAP evaluates the overall ranking quality over all correct matches, and mINP assesses the retrieval of the hardest positive sample.

Implementation Details. Our framework is implemented with PyTorch v1.13.1, and all experiments are conducted on a single NVIDIA GeForce RTX 3080 Ti GPU. We adopt the ViT-B/16, pretrained on ImageNet, as our backbone. The input images are resized to 256×128. During training, we employ standard data augmentation techniques, including random horizontal flipping, padding, and random erasing. The model is optimized using SGD with an initial learning rate of 8×10^{-3}, which is scheduled by cosine annealing to a minimum of 1.6×10^{-6}. The batch size is set to 128. We adopt a random identity sampler, where each mini-batch contains 32 identities with 4 images per identity. Regarding the hyperparameters in our loss functions and memory bank, the momentum coefficient α is set to 0.9. The balancing weights for the loss terms are set as $\beta = 1$, $\gamma = 1$, and $\xi = 0.01$. No test-time augmentation (TTA) or re-ranking post-processing is applied.

4.2 Performance

Performance on CARGO. As presented in Table 1, QVAM achieves state-of-the-art performance across all four evaluation protocols. Specifically, under the comprehensive ALL protocol, our method demonstrates superior performance, reaching 78.85% Rank-1, 71.02% mAP, and 60.25% mINP. This represents a significant improvement over the strong competitor SeCap with margins of +10.26% and +10.83% in Rank-1 and mAP, respectively. Even under the most challenging cross-view protocol (A↔G), QVAM still outperforms SeCap by +3.07% and +5.47%. Notably, our method also yields consistent gains on the same-view protocols, surpassing SeCap by +5.00% / +10.55% on A↔A and +4.46% / +7.20% on G↔G (Rank-1/mAP). These results indicate that QVAM effectively mitigates view-induced bias while preserving identity-discriminative cues for both cross-view and intra-view retrieval.

Performance on AG-ReID. Table 2 reports the results on AG-ReID. QVAM achieves the best performance under both cross-view protocols. On A→G, QVAM

Table 1: Performance comparison on CARGO. Four protocols: ALL, Cross-view (A↔G), Aerial-only (A↔A), and Ground-only (G↔G). **Bold** denotes the best result.

Method	Protocol 1: ALL			Protocol 2: A↔G			Protocol 3: A↔A			Protocol 4: G↔G		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
SBS [10]	50.32	43.09	29.76	31.25	29.00	18.71	67.50	49.73	29.32	72.31	62.99	48.24
PCB [32]	51.00	44.50	32.20	34.40	30.40	20.10	55.00	44.60	27.00	74.10	67.60	55.10
BoT [23]	54.81	46.49	32.40	36.25	32.56	21.46	65.00	49.79	29.82	77.68	66.47	51.34
MGN [34]	54.81	49.08	36.52	31.87	33.47	24.64	65.00	52.96	36.78	83.93	71.05	55.20
VV [17]	45.83	38.84	39.57	31.25	29.00	18.71	67.50	49.73	29.32	72.31	62.99	48.24
AGW [42]	60.26	53.44	40.22	43.57	40.90	29.39	67.50	56.48	40.40	81.25	71.66	58.09
ViT [4]	61.54	53.54	39.62	43.13	40.11	28.20	80.00	64.47	47.07	82.14	71.34	57.55
VDT [46]	64.10	55.20	41.13	48.12	42.76	29.95	82.50	66.83	50.22	82.14	71.59	58.39
DTST [37]	64.42	55.73	41.92	50.53	43.49	29.46	80.00	63.31	44.67	78.57	72.40	62.10
SeCap [36]	68.59	60.19	-	69.43	58.94	-	80.00	68.08	-	86.61	75.42	-
SD-ReID [38]	65.06	57.47	44.21	53.12	46.44	33.03	82.50	67.70	51.94	81.25	74.08	63.10
VIF [16]	65.71	57.46	44.12	51.25	44.55	31.20	82.50	66.98	51.44	83.93	74.19	62.30
QVAM (Ours)	**78.85**	**71.02**	**60.25**	**72.50**	**64.41**	**52.78**	**85.00**	**79.63**	**73.28**	**91.07**	**82.62**	**73.87**

Table 2: Performance comparison on AG-ReID. A/G denote aerial/ground cameras, respectively. **Bold** denotes the best result.

Method	A→G			G→A		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP
OSNet [53]	72.59	58.32	-	74.22	60.99	-
BoT [23]	70.01	55.47	-	71.20	58.83	-
SBS [10]	73.54	59.77	-	73.70	62.27	-
VV [17]	77.22	67.23	41.43	79.73	69.83	42.37
ViT [4]	81.28	72.38	-	82.64	73.35	-
TransReID [11]	81.80	73.10	-	83.40	74.60	-
VDT [46]	82.91	74.44	51.06	86.59	78.57	52.87
DTST [37]	83.48	74.51	49.86	84.72	76.05	50.04
SeCap [36]	84.03	76.16	-	87.01	78.34	-
SD-ReID [38]	85.16	75.40	51.35	85.97	77.02	50.42
VIF [16]	83.75	75.22	51.57	87.32	79.19	52.98
QVAM (Ours)	**85.53**	**76.81**	**53.04**	**88.46**	**80.52**	**55.12**

Table 3: Performance comparison on AG-ReIDv2. A/C/W denote aerial, CCTV, and wearable cameras, respectively. **Bold** denotes the best result.

Method	A→C		A→W		C→A		W→A	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
MGN [34]	82.09	70.17	88.14	78.66	84.21	72.41	84.06	73.73
BoT [23]	80.73	71.49	86.06	75.98	79.46	69.67	82.69	72.41
BoT+Attributes	81.43	72.19	86.66	76.68	80.15	70.37	83.29	73.11
SBS [10]	81.96	72.04	88.14	78.94	84.10	73.89	84.66	75.01
SBS+Attributes	82.56	72.74	88.74	79.64	84.80	74.59	85.26	75.71
ViT [4]	85.40	77.03	89.77	80.48	84.65	75.90	84.27	76.59
TransReID [11]	88.00	81.40	90.40	84.50	87.60	80.10	87.70	81.10
VDT [46]	86.46	79.13	90.00	82.21	86.14	78.12	85.26	78.52
SeCap [36]	88.12	80.84	**91.44**	84.01	**88.24**	79.99	87.56	80.15
SD-ReID [38]	-	80.57	-	84.09	-	78.90	-	80.07
QVAM (Ours)	**88.41**	**82.29**	90.49	**85.15**	87.24	**80.67**	**87.78**	**82.43**

attains 85.53% Rank-1, 76.81% mAP, and 53.04% mINP, outperforming the strong baseline VDT by +2.62%, +2.37%, and +1.98%, respectively. On G→A, QVAM reaches 88.46% Rank-1, 80.52% mAP, and 55.12% mINP, surpassing the previous best method VIF by +1.14%, +1.33%, and +2.14%. These results demonstrate the effectiveness of QVAM in mitigating aerial–ground viewpoint discrepancies and improving retrieval quality.

Performance on AG-ReIDv2. As summarized in Table 3, QVAM consistently improves over the previous best method SeCap in mAP across all four protocols. Specifically, QVAM achieves 88.41% Rank-1 / 82.29% mAP on A→C, improving SeCap by +0.29% Rank-1 and +1.45% mAP. On A→W, QVAM reaches 90.49% Rank-1 / 85.15% mAP, yielding a +1.14% mAP gain despite a slight Rank-1 drop (-0.95%). On C→A, QVAM attains 87.24% Rank-1 / 80.67% mAP, surpassing SeCap by +0.68% mAP with a small Rank-1 decrease (-1.00%). Notably, on the challenging W→A protocol, QVAM obtains 87.78% Rank-1 / 82.43% mAP, improving SeCap by +0.22% Rank-1 and +2.28% mAP. Overall, despite slight Rank-1 drops on A→W and C→A compared to SeCap, the consistent mAP gains indicate improved overall retrieval quality. We further quantify this behavior

Table 4: Performance stability of mAP (%) across random seeds on AG-ReIDv2.

Seed	A→C	A→W	C→A	W→A
1234	82.19	85.47	80.66	82.12
123891	82.29	85.15	80.67	82.43
328479	81.93	85.03	80.49	82.20
29304802	82.39	84.92	80.38	81.59
Mean ± Std	**82.20 ± 0.20**	**85.14 ± 0.24**	**80.55 ± 0.14**	**82.09 ± 0.36**

Table 5: Ablation study of each component on the CARGO dataset. VAD: View-aware Decoder; AFM: Adaptive Feature Modulation; CVPA: Cross-View Prototype Alignment. Best results are in **bold**.

Module				Protocol 1: ALL			Protocol 2: A↔G			Protocol 3: A↔A			Protocol 4: G↔G		
ViT	VAD	AFM	CVPA	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
✓				71.15	62.21	48.79	59.38	53.46	40.75	82.50	68.04	52.75	83.93	75.54	64.55
✓		✓		74.36	67.63	56.81	64.38	59.86	46.09	80.00	74.08	64.79	86.61	79.74	71.16
✓			✓	75.64	67.20	55.07	68.12	60.99	48.73	82.50	72.62	61.74	84.82	76.76	66.29
✓		✓	✓	74.68	70.00	**60.87**	66.87	62.98	52.60	82.50	78.59	72.69	85.71	80.23	72.79
✓	✓	✓		75.96	67.61	55.14	68.12	60.34	46.52	80.00	75.64	66.46	85.71	78.05	68.54
✓	✓	✓	✓	**78.85**	**71.02**	60.25	**72.50**	**64.41**	**52.78**	**85.00**	**79.63**	**73.28**	**91.07**	**82.62**	**73.87**

using mINP for the two protocols with slight Rank-1 drops. Compared with SeCap, QVAM improves mINP from 57.86 to 59.29 on A→W and from 58.29 to 60.28 on C→A, suggesting better hard-positive retrieval despite slight Rank-1 drops.

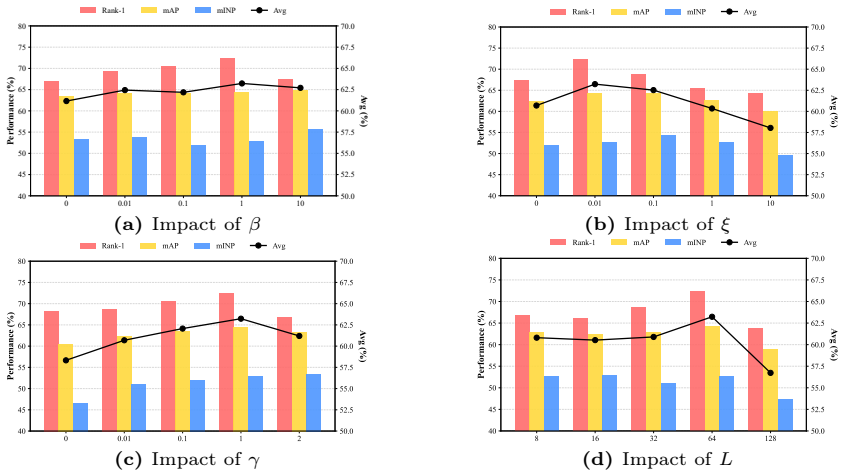
Stability across repeated runs. As shown in Table 4, QVAM shows low variance across 4 random seeds on all AG-ReIDv2 protocols (Std $\leq$ 0.36%). Its mean mAP also surpasses SeCap on all four protocols, suggesting that the observed improvements are stable rather than caused by random fluctuations.

4.3 Ablation Study

To validate the effectiveness of the proposed VAD, AFM and CVPA loss, we conduct comprehensive ablation studies on CARGO across four protocols.

Effectiveness of VAD. As shown in Table 5, introducing the VAD module significantly boosts performance. Specifically, under the most challenging A↔G protocol, the combination of VAD+AFM outperforms the AFM-only baseline by +3.74%, +0.48%, and +0.43% in Rank-1, mAP, and mINP, respectively. Furthermore, when integrating VAD into the AFM+CVPA setting (i.e., the full model vs. AFM+CVPA), the performance gains are even more pronounced, with improvements of +5.63%, +1.43%, and +0.18% in Rank-1, mAP, and mINP, respectively. These results verify that the proposed VAD successfully distills fine-grained viewpoint cues and provides crucial guidance for the subsequent adaptive feature modulation process.

Effectiveness of AFM. Comparing the AFM-only model with the ViT backbone, we observe substantial improvements of +5.00%, +6.40%, and +5.34% in



(a) Impact of β **(b)** Impact of ξ

(c) Impact of γ **(d)** Impact of L

Fig. 3: Parameter Sensitivity Analysis. Results on CARGO (A↔G). (a) Diversity loss weight β. (b) Mask regularization weight ξ. (c) CVPA loss weight γ. (d) Number of queries L.

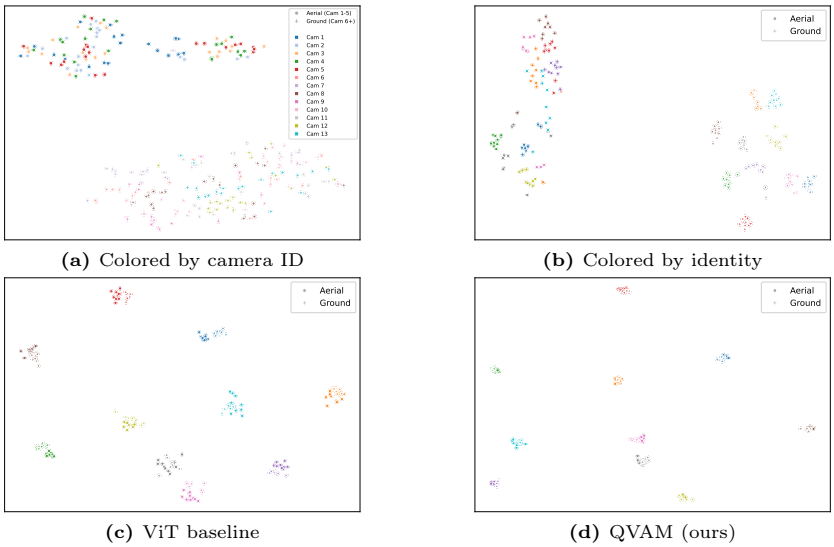
Rank-1, mAP, and mINP under the A↔G protocol. This demonstrates that even without explicit view queries, the AFM module effectively mitigates viewpoint discrepancies and adaptively modulates features to learn view-invariant identity representations by leveraging the `[CLS]` token.

Effectiveness of CVPA. The proposed CVPA loss is effective in aligning feature distributions. When applying CVPA directly to the ViT backbone (note that the loss is computed on the `[CLS]` token here), it yields clear gains of +8.74%, +7.53%, and +7.98% in Rank-1, mAP, and mINP under the A↔G protocol, respectively. When incorporated into the AFM and VAD+AFM architectures, CVPA further boosts Rank-1 accuracy by +2.49% and +4.38% under the A↔G protocol, respectively. Extensive experiments confirm that CVPA imposes robust constraints on view modulation and alignment, facilitating the generation of discriminative view-invariant representations.

4.4 Hyperparameter Analysis

To evaluate the effectiveness and sensitivity of our loss terms ($\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{reg}}$, $\mathcal{L}_{\text{CVPA}}$) and architectural choices, we conduct extensive experiments on the CARGO dataset under the challenging A↔G protocol.

Effect of Diversity Loss Weight (β). Fig. 3(a) shows that enabling $\mathcal{L}_{\text{div}}$ ($\beta > 0$) consistently improves performance over the baseline ($\beta = 0$). This suggests that promoting query diversity helps VAD capture complementary viewpoint cues for subsequent modulation. Performance increases with β up to a moderate range, but degrades when β becomes too large: overly strong diversity regularization may encourage queries to separate by focusing on noisy or



(a) Colored by camera ID **(b)** Colored by identity

(c) ViT baseline **(d)** QVAM (ours)

Fig. 4: *t*-SNE visualization of view and identity embeddings. (a-b) Explicit view features colored by camera ID and identity, respectively (markers denote aerial vs. ground). (c-d) Identity embeddings learned by the ViT baseline and our QVAM.

irrelevant patterns, whereas too small β may be insufficient to prevent query collapse.

Effect of Mask Regularization Weight (ξ). As shown in Fig. 3(b), mild regularization ($\xi = 0.01$ or 0.1) outperforms the unregularized setting ($\xi = 0$). This indicates that $\mathcal{L}_{\text{reg}}$ helps the modulation mask suppress viewpoint-related interference and encourages view-invariant representations. However, performance drops noticeably when $\xi > 0.1$. View-related channels are not necessarily pure nuisance signals in AGPReID, since they may still retain identity-discriminative information. Overly strong regularization may suppress useful identity cues entangled with viewpoint-sensitive channels.

Effect of CVPA Loss Weight (γ). Fig. 3(c) shows that enabling the CVPA loss ($\gamma > 0$) yields clear gains over $\gamma = 0$, supporting its crucial role in reducing aerial–ground misalignment through distribution-level alignment. Performance peaks around $\gamma = 1$ and degrades for larger γ, likely because overly strong prototype matching amplifies prototype noise (e.g., under limited or imbalanced samples) and harms generalization.

Effect of Number of Queries (L). Fig. 3(d) demonstrates that increasing L from 8 to 64 improves performance, suggesting that sufficient queries are beneficial for covering diverse local viewpoint details. Further increasing L to 128 leads to a sharp drop, which we attribute to redundancy: too many queries may attend to background noise or irrelevant patterns, interfering with the modulation process.

Fig. 5: Visualization of VAD query-to-patch attention. Average attention maps from different VAD queries to patch tokens in aerial and ground views, illustrating that the learned queries consistently focus on human body regions and highlight view-sensitive areas induced by viewpoint changes.

4.5 Visualization

To better understand the challenges of AGPReID and the effectiveness of QVAM, we provide qualitative analyses using t-SNE visualizations and attention heatmaps on CARGO under the ALL protocol.

Analysis of View Features (Motivation). In Fig. 4(a,b), we visualize explicit view-specific features generated by a ViT baseline augmented with a view token and trained with cross-entropy and triplet losses using binary view labels. Fig. 4(a) clearly shows that aerial samples do not form a single compact cluster. This suggests that continuous altitude and depression-angle changes create significant intra-view diversity, which coarse binary supervision cannot fully capture. Fig. 4(b) further reveals identity-wise local grouping within the same view, indicating that view-specific features still encode identity-discriminative information rather than purely nuisance factors. This explains why overly strong regularization may suppress useful identity cues, motivating our choice of query-guided adaptive modulation instead of rigid disentanglement.

Analysis of Feature Alignment (Effectiveness). Fig. 4(c,d) compares the final identity representations of the ViT baseline and QVAM. The Baseline suffers from severe domain separation, where aerial and ground samples of the same identity are far apart. In contrast, QVAM successfully aligns the distributions, pulling positive pairs from different views into compact clusters. This demonstrates that by leveraging fine-grained viewpoint cues and adaptive feature modulation mechanism, our model effectively improves aerial–ground alignment and learns more view-invariant representations.

Visualization of Attention Map. We further visualize the average attention maps between learnable queries and patch tokens for four different identities

Fig. 6: Visualization of query-to-patch attention under increasing viewpoint changes. The learnable queries adaptively adjust their attention distributions to capture fine-grained viewpoint cues of the same identity under increasing viewpoint changes.

across aerial and ground views. As shown in Fig. 5, despite the absence of explicit spatial supervision, queries consistently attend to human body regions rather than background clutter, indicating that view-aware queries encode semantically meaningful human-centric contexts. Furthermore, by comparing the attention maps across views, we observe that the focus intensity varies adaptively: the queries capture not only shared identity cues but also view-specific variations, particularly in the lower-body regions, which are highly sensitive to altitude changes. This suggests that view-aware queries effectively distill fine-grained viewpoint cues, providing informative guidance for subsequent view-adaptive mask prediction to perform adaptive feature modulation.

Quantitative Proxy for Viewpoint Modeling. We use a balanced camera-ID linear probe on CARGO ALL as a quantitative proxy. With 3,133 uniformly sampled images per camera, a logistic-regression classifier on the learned queries achieves 89.48% accuracy, versus 7.68% for static queries (near the 1/13 random-guess level). Although camera ID is only a coarse proxy, this result and the same-identity attention visualization (shown in Fig. 6) under increasing viewpoint changes support that the learned queries capture viewpoint-sensitive cues rather than merely serving as generic conditioning vectors.

5 Conclusion

In this work, we propose the QVAM framework to address the extreme viewpoint gap in AGPReID. We observe that coarse binary aerial/ground labels are overly simplified for modeling continuous aerial viewpoint variations. To this end, QVAM distills fine-grained viewpoint cues via learnable view queries and performs identity-preserving adaptive modulation of image-level representations, while a Cross-View Prototype Alignment (CVPA) loss further enforces aerial–ground feature alignment. Extensive experiments on multiple benchmarks demonstrate consistent improvements over prior methods, supporting the effectiveness of QVAM under severe cross-view domain gaps.

References

1. Chen, T., Chen, C., Zhou, Y., Hu, X.: Global-local unified cross-view enhancing framework for person re-identification. In: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–9. IEEE (2024)
2. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 587–597 (2021)
3. Choi, S., Kim, T., Jeong, M., Park, H., Kim, C.: Meta batch-instance normalization for generalizable person re-identification. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 3425–3435 (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Ge, Y., Zhu, F., Chen, D., Zhao, R., et al.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Advances in neural information processing systems **33**, 11309–11321 (2020)
6. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision. pp. 262–275. Springer (2008)
7. Han, K., Huang, Y., Gong, S., Wang, L., Tan, T.: 3d shape temporal aggregation for video-based clothing-change person re-identification. In: Proceedings of the Asian Conference on Computer Vision. pp. 2371–2387 (2022)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
9. 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. pp. 770–778 (2016)
10. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9664–9667 (2023)
11. 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. pp. 15013–15022 (2021)
12. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017)
13. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
14. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3143–3152 (2020)
15. Khaldi, K., Nguyen, V.D., Mantini, P., Shah, S.: Unsupervised person re-identification in aerial imagery. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 260–269 (2024)

16. Khalid, W., Liu, B., Li, X., Waqas, M., Afgan, M.S.: 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)
17. Kumar, R., Weill, E., Aghdasi, F., Sriram, P.: A strong and efficient baseline for vehicle re-identification using deep triplet embedding. Journal of Artificial Intelligence and Soft Computing Research **10**(1), 27–45 (2020)
18. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 4610–4617 (2020)
19. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16266–16275 (2021)
20. Li, Y.J., Weng, X., Kitani, K.M.: Learning shape representations for person re-identification under clothing change. In: Proceedings of the IEEE/CVF Winter conference on applications of computer vision. pp. 2432–2441 (2021)
21. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2197–2206 (2015)
22. Liu, F., Ashbaugh, R., Chimitt, N., Hassan, N., Hassani, A., Jaiswal, A., Kim, M., Mao, Z., Perry, C., Ren, Z., et al.: Farsight: A physics-driven whole-body biometric system at large distance and altitude. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6227–6236 (2024)
23. 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 (CVPR) Workshops (June 2019)
24. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 542–551 (2019)
25. Nguyen, H., Nguyen, K., Sridharan, S., Fookes, C.: Aerial-ground person re-id. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2585–2590. IEEE (2023)
26. 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 **19**, 2896–2908 (2024)
27. Nguyen, K., Fookes, C., Sridharan, S., Liu, F., Liu, X., Ross, A., Michalski, D., Nguyen, H., Deb, D., Kothari, M., et al.: Ag-reid 2023: Aerial-ground person re-identification challenge results. In: 2023 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2023)
28. Nguyen, K., Fookes, C., Sridharan, S., Nguyen, H., Liu, F., Liu, X., Ross, A., Endrei, T., DeAndres-Tame, I., Tolosana, R., et al.: Ag-vpreid 2025: Aerial-ground video-based person re-identification challenge results. In: 2025 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2025)
29. Pang, Z., Wang, J., Zhao, L., Wang, C.: Identity-clothing similarity modeling for unsupervised clothing change person re-identification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19251–19260 (2025)
30. Sarker, P.K., Zhao, Q., Uddin, M.K.: Transformer-based person re-identification: a comprehensive review. IEEE Transactions on Intelligent Vehicles **9**(7), 5222–5239 (2024)

31. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 719–728 (2019)
32. Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q., Wang, S.: Learning part-based convolutional features for person re-identification. IEEE transactions on pattern analysis and machine intelligence **43**(3), 902–917 (2019)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. 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. pp. 274–282 (2018)
35. Wang, L., Zhang, Q., Qiu, J., Lai, J.: Rotation exploration transformer for aerial person re-identification. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
36. 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 Computer Vision and Pattern Recognition Conference. pp. 22119–22128 (2025)
37. Wang, Y., Pishgar, M.: Dynamic token selection for aerial-ground person re-identification. arXiv preprint arXiv:2412.00433 (2024)
38. Wang, Y., Hu, X., Wang, L., Zhang, P., Lu, H.: Sd-reid: View-aware stable diffusion for aerial-ground person re-identification. arXiv preprint arXiv:2504.09549 (2025)
39. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
40. Wu, L., Hong, R., Wang, Y., Wang, M.: Cross-entropy adversarial view adaptation for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **30**(7), 2081–2092 (2019)
41. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: European conference on computer vision. pp. 1–16. Springer (2014)
42. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
43. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 994–1002 (2017)
44. Zhang, C., Wu, L., Wang, Y.: Crossing generative adversarial networks for cross-view person re-identification. Neurocomputing **340**, 259–269 (2019)
45. Zhang, G., Yang, Y., Zheng, Y., Martin, G., Wang, R.: Mask-aware hierarchical aggregation transformer for occluded person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **35**(6), 5821–5832 (2025)
46. 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. pp. 22000–22009 (2024)
47. Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., Zhang, Y.: Person re-identification in aerial imagery. IEEE Transactions on Multimedia **23**, 281–291 (2020)

48. Zhang, X., Yan, Y., Xue, J.H., Hua, Y., Wang, H.: Semantic-aware occlusion-robust network for occluded person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **31**(7), 2764–2778 (2020)
49. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
50. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2138–2147 (2019)
51. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 598–607 (2019)
52. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
53. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5056–5069 (2021)
54. Zhu, K., Guo, H., Zhang, S., Wang, Y., Liu, J., Wang, J., Tang, M.: Aaformer: Auto-aligned transformer for person re-identification. IEEE transactions on neural networks and learning systems (2023)
55. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: 2018 IEEE international conference on multimedia and expo (ICME). pp. 1–6. IEEE (2018)