# Hierarchical Hyperbolic Representation for Aerial-Ground Person Re-Identification

Qiwei Yang✉

Dalian University of Technology, Dalian, China
✉ Corresponding author: dutyqw@mail.dlut.edu.cn

**Abstract.** Aerial-Ground Person Re-IDentification (AG-ReID) aims to retrieve the same person across heterogeneous aerial and ground camera platforms. Although great progress, existing methods remain suboptimal due to the direct feature alignment across views, overlooking view-specific cues. To address this issue, we propose a novel Hierarchical Hyperbolic Representation (HiHR) framework for AG-ReID. More specifically, we first extract multi-granularity features based on pre-trained visual-text encoders. Then, we propose a Text-guided Multi-granularity Fusion (TMF) to fuse multi-granularity features and enhance the representation ability of identity features. Furthermore, we introduce the Hierarchical Hyperbolic Learning (HHL) to construct a hierarchical feature structure in a hyperbolic space. This hierarchy includes a coarse level that ensures identity separability and cross-view consistency, and a fine level that preserves view-specific discriminative cues. A a result, our proposed framework can effectively aggregate view-invariant and view-specific discriminative features for AG-ReID. Extensive experiments on four AG-ReID benchmarks demonstrate the effectiveness of our framework. The source code is available at https://github.com/YangQiWei3/HiHR.

**Keywords:** Aerial-Ground Person Re-Identification · Hyperbolic Representation Learning · Hierarchical Metric Learning

## 1 Introduction

Person Re-IDentification (ReID) aims to retrieve the same person across non-overlapping cameras, which serves as a fundamental capability for intelligent surveillance systems [11, 12, 26]. Despite great progress, existing methods [14, 23, 29] mainly focus on homogeneous camera networks, limiting their applicability to real-world scenarios. In practice, real-world deployments increasingly integrate heterogeneous camera platforms, where ground cameras provide detailed close-range observations, while aerial sensors offer wide-area coverage. This motivates Aerial-Ground Person Re-IDentification (AG-ReID), which aims to retrieve persons across heterogeneous platforms under systematic viewpoint discrepancies. Such discrepancies introduce substantial appearance gaps, hindering discriminative representation learning. However, as shown in Fig. 1(a), the discriminative representation should be view-agnostic and allow cross-view discrepancies.

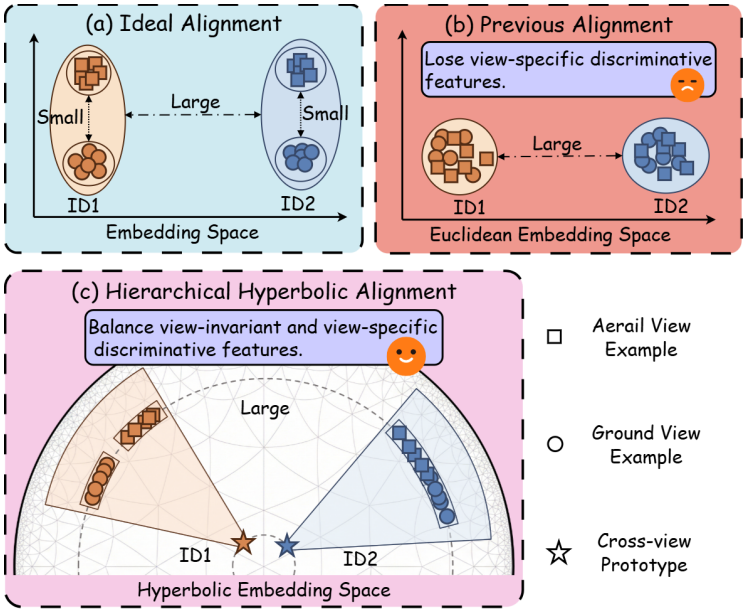


**Fig. 1:** Illustration of our motivations and proposed framework. (a) The ideal alignment clusters examples of the same identity nearby while different identities far apart. (b) Previous alignments in Euclidean space align cross-view samples directly and may suppress view-specific discriminative cues. (c) Our hierarchical hyperbolic alignment keeps a hierarchical separability and view-specific cues.

To this end, previous methods [5, 13, 22, 31, 32] directly align samples across views, as shown in Fig. 1(b). Although effective, they emphasize view-invariant features while overlooking view-specific cues. Moreover, many AG-ReID methods [3, 5, 7, 13, 33] rely solely on global representations, which capture high-level semantics. They may attenuate mid-level structural cues and fine-grained local patterns for reliable retrieval.

To address the aforementioned issues, we propose a novel Hierarchical Hyperbolic Representation (HiHR) framework for AG-ReID. More specifically, inspired by CLIP-ReID [14], we first introduce text prompts to extract multi-granularity features based on pre-trained visual-text encoders. Then, we propose a Text-guided Multi-granularity Fusion (TMF) to fuse the multi-granularity features and enhance the representation ability of identity features. Furthermore, we introduce the Hierarchical Hyperbolic Learning (HHL) to construct a hierarchical feature structure in a hyperbolic space, as illustrated in Fig. 1(c). It includes a coarse level for identity separability and cross-view consistency, together with a fine level for preserving view-specific discriminative cues. The entailment-cone regularization is further introduced to maintain hierarchical consistency by propagating identity-consistent separation from the coarse level to the fine level. A a result, our proposed framework can effectively aggregate view-invariant and

view-specific discriminative features for AG-ReID. Extensive experiments on four AG-ReID benchmarks demonstrate the effectiveness of our framework.

In summary, our contributions are as follows:

- We propose a novel Hierarchical Hyperbolic Representation (HiHR) framework for AG-ReID. To our best knowledge, it is the first work to utilize hyperbolic representations for cross-view person retrieval.
- We propose a Text-guided Multi-granularity Fusion (TMF) to fuse view-invariant and view-specific discriminative features in multi-granularity.
- We propose the Hierarchical Hyperbolic Learning (HHL) to model coarse-to-fine representation structures of cross-view and view-specific features.
- Extensive experiments on four benchmarks demonstrate that our proposed framework achieves competitive or state-of-the-art performance.

## 2 Related Work

### 2.1 Aerial-Ground Person Re-Identification

AG-ReID extends conventional ground-to-ground ReID [6, 24, 27, 28, 30] to heterogeneous aerial-ground camera networks. Early works [16, 17, 22, 33] mainly establish benchmarks and baselines. Building on these foundations, recent methods [13, 15, 36] propose view-aware models to mitigate the aerial-ground discrepancies. Technically, Zhang *et al.* [33] propose a view-decoupled Transformer to separate view-related and view-unrelated information. Li *et al.* [13] develop a geometric-and-semantic alignment network to compensate cross-view distortion and emphasize visibility-consistent regions. Mao *et al.* [15] introduce learnable text tokens to enhance cross-view feature consistency. Huang *et al.* [5] propose a perspective-driven prototype alignment to stabilize cross-view matching. Despite promising results, they directly align samples across views, overlooking view-specific discriminative information. This limitation motivates our framework, which aims to maintain cross-view identity consistency while preserving view-specific discriminative cues.

### 2.2 Hyperbolic Representation Learning for ReID

Hyperbolic representation learning provides an effective inductive bias for hierarchical or tree-structured data, by leveraging the exponential volume growth induced by negative curvatures [1, 10]. This property has motivated a growing body of work that leverages hyperbolic geometry to better preserve semantic granularity, improve metric learning and enhance representation separability. However, in person ReID, hyperbolic representation learning has not been fully explored. Early attempts mainly adopt hyperbolic geometry as a drop-in replacement of the embedding space. For example, Khrulkov *et al.* [8] append hyperbolic layers to standard visual encoders and project image features into a Poincaré ball, where the similarity is computed with a hyperbolic distance. Fang *et al.* [2] develop positive definite kernels for hyperbolic spaces and evaluate them

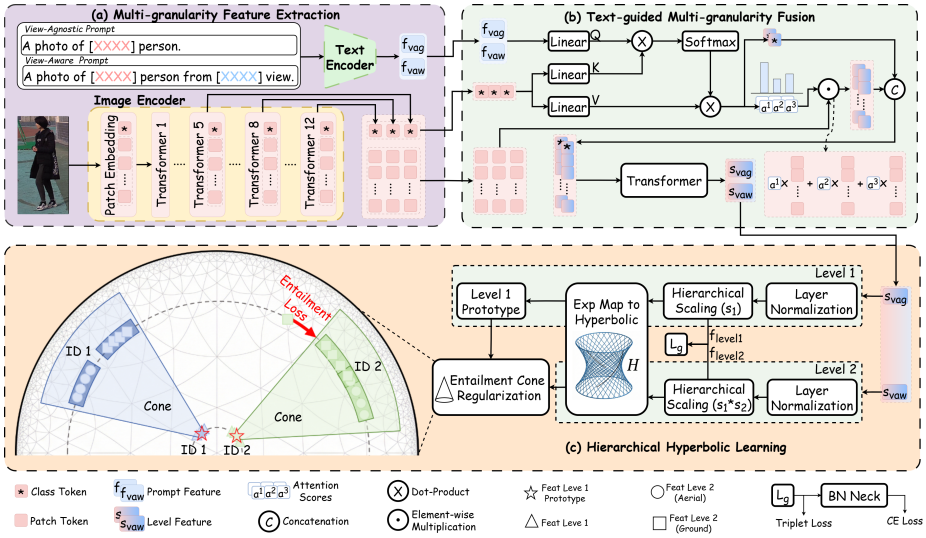


**Fig. 2:** Illustration of our proposed framework. It includes three key key modules, i.e., Multi-granularity Feature Extraction (MFE), Text-guided Multi-granularity Fusion (TMF) and Hierarchical Hyperbolic Learning (HHL).

on person ReID. Although effective, existing methods usually treat the hyperbolic space as a single-level embedding or similarity metric. This makes them difficult to capture persons' complex appearance variations caused by viewpoint discrepancies in AG-ReID. To address this limitation, we propose HHL to embed multi-granularity representations into a single Lorentz hyperbolic manifold [1], thereby modeling the coarse-to-fine feature hierarchy for AG-ReID.

## 3 Methodology

In this section, we introduce the proposed Hierarchical Hyperbolic Representation (HiHR) framework for AG-ReID. As shown in Fig. 2, it includes three key key modules, i.e., Multi-granularity Feature Extraction (MFE), Text-guided Multi-granularity Fusion (TMF) and Hierarchical Hyperbolic Learning (HHL). Details are described as follows.

### 3.1 Multi-granularity Feature Extraction

In practice, many AG-ReID methods rely solely on global representations, which capture high-level semantics. However, they may attenuate mid-level structural cues and fine-grained local patterns for reliable retrieval. To improve the feature representation ability, we first introduce the Multi-granularity Feature Extraction (MFE). More specifically, inspired by CLIP-ReID [14], we extract multi-granularity features based on pre-trained visual-text encoders, such as CLIP. For visual features, we utilize the output of different Transformer layers from

the image encoder $E_v$. For an input image $\mathbf{I}$, the token sequence at layer $l$ is denoted as

$$\mathbf{Z}^{(l)} = [\mathbf{c}^{(l)}; \mathbf{X}^{(l)}] \in \mathbb{R}^{(1+N)\times D}, \tag{1}$$

where $\mathbf{c}^{(l)}$ and $\mathbf{X}^{(l)}$ denote the class token and patch tokens, respectively. $N$ is the number of patch tokens. $D$ is the token dimension.

For text features, we introduce dual-level prompts to extract view-agnostic and view-aware information. As shown in Fig. 2(a), the view-agnostic prompt $\mathbf{p}_{vag}$ is defined as "*A photo of* ⌈shared⌋ *person.*". It emphasizes identity-consistent person evidence. The corresponding embedding is obtained by the text encoder $E_t$ as follows:

$$\mathbf{f}_{vag} = E_t(\mathbf{p}_{vag}) \in \mathbb{R}^D. \tag{2}$$

In parallel, the view-aware prompt $\mathbf{p}_{vaw}$ is defined as "*A photo of* ⌈shared⌋ *person from* ⌈view-private⌋ *view.*". It explicitly contains viewpoint information and captures view-specific discriminative cues within the corresponding view. The corresponding embedding is obtained as follows:

$$\mathbf{f}_{vaw} = E_t(\mathbf{p}_{vaw}) \in \mathbb{R}^D. \tag{3}$$

### 3.2 Text-guided Multi-granularity Fusion

Technically, one can concatenate the above multi-granularity features, and perform person retrieval. However, there are modality gaps in the visual features and text features, which may lead to inferior performance. To address this issue, we propose a Text-guided Multi-granularity Fusion (TMF) to fuse multi-granularity features and enhance the representation ability of identity features.

To integrate complementary information from different granularities, we perform a cross-attention by using the text features as semantic queries and the class tokens from different Transformer layers as keys and values, as follows:

$$\boldsymbol{\alpha}_k = \mathrm{softmax}\left(\frac{\mathbf{Q}_k\mathbf{K}^\top}{\sqrt{D}}\right), \quad \mathbf{t}_k = \boldsymbol{\alpha}_k\mathbf{V}, \quad k \in \{vag, vaw\}, \tag{4}$$

where $(\mathbf{Q}_k, \mathbf{K}, \mathbf{V})$ are linear projections of $(\mathbf{f}_k, \mathbf{C}, \mathbf{C})$. $\mathbf{C} = [\mathbf{c}^{(l_1)}; \mathbf{c}^{(l_2)}; \ldots; \mathbf{c}^{(l_M)}]$. $M$ is the total number of selected layers. Thus, we can obtain the view-agnostic fused feature $\mathbf{t}_{vag}$ and the view-aware fused feature $\mathbf{t}_{vaw}$, respectively.

Meanwhile, fine-grained discriminative cues often resides in patch tokens. Thus, we reuse $\boldsymbol{\alpha}_k$ to aggregate multi-layer patch tokens as follows:

$$\mathbf{X}_k = \sum_{m=1}^{M} \alpha_k^{(m)}\, \mathbf{X}^{(l_m)}. \tag{5}$$

We then fuse $\mathbf{X}_k$ with $\mathbf{t}_k$ through a standrd Transformer layer:

$$\mathbf{S}_k = \mathrm{Trans}\big([\mathbf{t}_k; \mathbf{X}_k]\big). \tag{6}$$

We take the first token $\mathbf{s}_k \in \mathbb{R}^D$ in $\mathbf{S}_k$ as the output representation. As observed, our TMF can adaptive fuse multi-granularity features. The fused feature provides a better representation for robust cross-view retrieval and serve as the inputs to the subsequent hierarchical hyperbolic learning.

### 3.3 Hierarchical Hyperbolic Learning

In fact, AG-ReID needs to maintain cross-view identity consistency while preserving view-specific discriminative cues. To this end, we propose Hierarchical Hyperbolic Learning (HHL) to construct a hierarchical feature structure in a hyperbolic space. As shown in Fig. 1(c), it includes a coarse level that ensures identity separability and cross-view consistency, and a fine level that preserves view-specific discriminative cues.

**Hyperbolic Representation.** A hyperbolic space is a smooth Riemannian manifold with a constant negative sectional curvature $-c$ ($c$>0) [1,10]. To obtain the hyperbolic representation, we define the following formulas. The exponential map $\exp_{\mathbf{x}}^{c}(\cdot)$ projects tangent vectors at $\mathbf{x}$ to a Lorentz space [1]. The logarithmic map $\log_{\mathbf{y}}^{c}(\cdot)$ projects points on the Lorentz space to the tangent space at $\mathbf{y}$. Third, for each $\mathbf{x}$, hyperbolic entailment cones $\omega(\mathbf{x})$ specify a region whose points $\mathbf{y} \in \omega(\mathbf{x})$ are treated as child concepts of $\mathbf{x}$. The half-aperture is obtained by

$$\omega(\mathbf{x}) = \sin^{-1}\left(\frac{2k}{\sqrt{c}\,\|\mathbf{x}_{\text{space}}\|}\right), \tag{7}$$

where $k = 0.1$ is a constant and $\mathbf{x}_{\text{space}}$ denote the spatial coordinates of $\mathbf{x}$. To ensure the child embeddings $\mathbf{y}$ within the parent cones $\omega(\mathbf{x})$, one can define the entailment loss as

$$\mathcal{L}_{\text{e}}(\mathbf{x}, \mathbf{y}) = \max\big(0, \phi(\mathbf{x}, \mathbf{y}) - \omega(\mathbf{x})\big), \tag{8}$$

where $\phi(\mathbf{x}, \mathbf{y})$ is the angle from $\mathbf{x}$ to $\mathbf{y}$.

For HHL, we take the two complementary embeddings from TMF as the parent and child representations. To reflect the coarse-to-fine hierarchy, we normalize the two representations and impose learnable scale separations:

$$\mathbf{g}^{p} = s_1 \frac{\mathbf{s}_{vag}}{\|\mathbf{s}_{vag}\|_2}, \qquad \mathbf{g}^{c} = s_1 s_2 \frac{\mathbf{s}_{vaw}}{\|\mathbf{s}_{vaw}\|_2}, \qquad s_1 > 0, s_2 > 1, \tag{9}$$

where $s_1$ controls the global embedding scale and $s_2$ pushes child embeddings farther from the origin than parent embeddings. This scale separation provides a stable hierarchical initialization and reduces interference between the two levels.

**Entailment Regularization for Hierarchical Structure.** Scale separation provides an initial coarse-to-fine hierarchy, but child clusters may still drift and overlap across identities. To impose identity-aware hierarchical constraints, we first compute parent prototypes from the scaled parent features in Euclidean space. For each identity $y$ in a batch, the parent prototype $\boldsymbol{\pi}_y^{p}$ is defined as

$$\boldsymbol{\pi}_y^{p} = \text{Mean}\big(\{\mathbf{g}^{p} \mid y_i = y\}\big), \tag{10}$$

where Mean$(\cdot)$ denotes the mean operation. Each prototype serves as the center of an identity-specific parent cone. We then project $\mathbf{g}^{p}$, $\mathbf{g}^{c}$ and $\boldsymbol{\pi}_y^{p}$ to the Lorentz hyperbolic space using the exponential map $\exp_{\mathbf{x}}^{c}(\cdot)$ at the origin $\mathbf{o} = [\frac{1}{\sqrt{c}}, \mathbf{0}]$:

$$\mathbf{z}^{p} = \exp_{\mathbf{o}}^{c}(\mathbf{g}^{p}), \qquad \mathbf{z}^{c} = \exp_{\mathbf{o}}^{c}(\mathbf{g}^{c}), \qquad \mathbf{z}^{\boldsymbol{\pi}} = \exp_{\mathbf{o}}^{c}(\boldsymbol{\pi}_y^{p}). \tag{11}$$

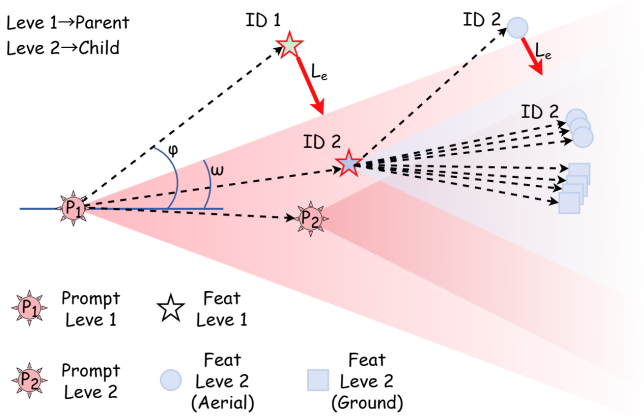


**Fig. 3:** Illustration of entailment regularizations for hierarchical consistency in HHL. Four entailment constraints are imposed: parent prototypes to child embeddings ($\mathbf{z}^{\boldsymbol{\pi}} \rightarrow \mathbf{z}^c$), parent prompts to child prompts ($\mathbf{z}_P^p \rightarrow \mathbf{z}_P^c$), parent prompts to parent visual features ($\mathbf{z}_P^p \rightarrow \mathbf{z}^p$), and child prompts to child visual features ($\mathbf{z}_P^c \rightarrow \mathbf{z}^c$). These constraints jointly preserve hierarchical consistency, align vision-language semantics, and limit child-cluster drift while retaining view-specific discriminative cues.

Given the prototype $\mathbf{z}^{\boldsymbol{\pi}}$, its entailment cone is centered at $\mathbf{z}^{\boldsymbol{\pi}}$, with aperture $\omega(\mathbf{z}^{\boldsymbol{\pi}})$ determined by the prototype distance to the origin. The child embedding $\mathbf{z}^c$ is regularized to lie within the corresponding identity-specific cone:

$$\mathcal{L}_{\mathrm{e}}(\mathbf{z}^{\boldsymbol{\pi}}, \mathbf{z}^c) = \max\big(0, \phi(\mathbf{z}^{\boldsymbol{\pi}}, \mathbf{z}^c) - \omega(\mathbf{z}^{\boldsymbol{\pi}})\big). \tag{12}$$

This constraint confines child-level refinement to the parent-induced feasible region, preserving view-specific discriminative cues while mitigating uncontrolled drift and cross-identity confusion.

**Entailment Regularization for Visual-Prompt Consistency.** Beyond the prototype-to-child visual constraint, we further align the prompt hierarchy with the visual hierarchy. The view-agnostic feature $\mathbf{f}_{vag}$ provides broader semantic guidance than the view-aware feature $\mathbf{f}_{vaw}$. This is because $\mathbf{f}_{vag}$ encodes general identity-related semantics shared across views, whereas $\mathbf{f}_{vaw}$ incorporates view-conditioned semantic details. Accordingly, we treat $\mathbf{f}_{vag}$ as the parent prompt and regularize it to entail the child prompt $\mathbf{f}_{vaw}$. Following the same scale separation, the two features are projected to the hyperbolic space as

$$\mathbf{z}_P^p = \exp_{\mathbf{o}}^c(s_1 \frac{\mathbf{f}_{vag}}{\|\mathbf{f}_{vag}\|_2}), \qquad \mathbf{z}_P^c = \exp_{\mathbf{o}}^c(s_1 s_2 \frac{\mathbf{f}_{vaw]}}{\|\mathbf{f}_{vaw}\|_2}), \tag{13}$$

and the prompt-level hierarchy is enforced by $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}_P^c)$.

Since text generally provides broader context than images [19], we further constrain visual embeddings to be entailed by their same-level prompt embeddings. This yields the parent-level and child-level visual-semantic constraints $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}^p)$ and $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^c, \mathbf{z}^c)$, respectively. Together with the prototype-to-child

constraint, these four entailment constraints form the hierarchical structure, as shown in Fig. 3. They preserve the semantic consistency between prompts and visual representations.

**Hierarchical Supervision.** In addition to the entailment regularizations, we employ standard ReID objectives to supervise the two hierarchy levels. Given an input feature $\mathcal{F}$, the objective combines label-smoothing cross-entropy loss $\mathcal{L}_{\text{CE}}$ [21] and triplet loss $\mathcal{L}_{\text{Tri}}$ [4]:

$$\mathcal{L}_g(\mathcal{F}) = \lambda_g \mathcal{L}_{\text{CE}}(\mathcal{F}) + \mathcal{L}_{\text{Tri}}(\mathcal{F}), \tag{14}$$

At the parent level, $\mathcal{L}_g(\mathbf{g}^p)$ is applied to learn view-invariant identity semantics. This supervision compacts same-identity samples across aerial and ground views, thereby strengthening coarse-level cross-view consistency. At the child level, $\mathcal{L}_{g,\text{intra}}(\mathbf{g}^c)$ is applied within each view to avoid over-suppressing view-dependent variations. Specifically, positives and negatives in $\mathcal{L}_{\text{Tri,intra}}(\mathbf{g}^c)$ are sampled from the same view as the anchor. This intra-view supervision preserves view-specific discriminative cues and refines the local identity structure. Together, the parent and child objectives impose cross-view compactness at the coarse level and view-aware compactness at the fine level.

### 3.4 Training and Inference

**Overall Objective.** The final training objective is a weighted combination of the above losses and entailment regularizations:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_g(\mathbf{g}^p) + \mathcal{L}_{g,\text{intra}}(\mathbf{g}^c) \\ + \lambda_e \Big( \mathcal{L}_{\text{e}}(\mathbf{z}^{\boldsymbol{\pi}}, \mathbf{z}^c) + \mathcal{L}_{\text{e}}(\mathbf{z}_P^p, \mathbf{z}_P^c) + \mathcal{L}_{\text{e}}(\mathbf{z}_P^p, \mathbf{z}^p) + \mathcal{L}_{\text{e}}(\mathbf{z}_P^c, \mathbf{z}^c) \Big),\end{aligned} \tag{15}$$

where $\lambda_e$ is a hyperparameter that balances the objectives.

**Inference.** During testing, we map $\mathbf{z}^p$ and $\mathbf{z}^c$ back to Euclidean space using the logarithmic map $\log_{\mathbf{y}}^c(\cdot)$. Then, we reduce them by $\frac{1}{s_1}$ and $\frac{1}{s_1 s_2}$, respectively, and concatenate them to obtain the final retrieval feature.

## 4 Experiments

In this section, we first introduce the datasets, evaluation protocols and implementation details. Then, we compare our method with state-of-the-arts. Finally, ablation studies are conducted to verify the contribution of each component.

### 4.1 Datasets and Evaluation Protocols.

We evaluate our framework on four AG-ReID benchmarks. To be specific, the AG-ReID [18] includes 21,983 images of 388 identities captured by one ground camera and one aerial camera. The AG-ReID v2 [17] comprises 100,502 images of 1,615 identities collected from three platforms, including UAV, wearable cameras

**Table 1: Performance comparison on AG-ReID and AG-ReID v2.** Best results are in **bold**.

| Method | AG-ReID | | | | AG-ReID v2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A→G | | G→A | | A→C | | C→A | | A→W | | W→A | |
| | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| Explain [16] | 72.38 | 81.28 | 73.35 | 82.64 | 79.00 | 87.70 | 78.24 | 87.35 | 83.14 | **93.67** | 79.08 | 87.73 |
| VDT [33] | 74.44 | 82.91 | 78.57 | 86.59 | 79.13 | 86.46 | 78.12 | 86.14 | 82.21 | 90.00 | 78.52 | 85.26 |
| GSAlign [13] | 75.01 | 83.75 | 77.73 | 84.10 | 81.38 | 87.86 | 81.05 | 88.02 | 83.98 | 90.63 | 80.90 | 87.31 |
| SD-ReID [26] | 75.40 | 85.16 | 77.02 | 85.57 | 80.61 | 87.04 | 79.24 | 86.74 | 84.06 | 90.86 | 80.12 | 86.79 |
| SAS-VPReID [31] | 75.98 | 84.24 | 77.28 | 86.17 | 76.39 | 83.40 | 78.47 | 85.26 | 82.10 | 89.45 | 78.47 | 85.73 |
| SeCap [22] | 76.16 | 84.03 | 78.34 | 87.01 | 80.84 | 88.12 | 79.99 | 88.24 | 84.01 | 91.44 | 80.15 | 87.56 |
| LATex [32] | 77.67 | 85.26 | 81.15 | **89.40** | 83.50 | **89.13** | 82.85 | **89.01** | 86.35 | 91.35 | 83.30 | 89.32 |
| **HiHR** | **79.21** | **86.06** | **81.28** | 87.94 | **84.04** | 88.84 | **83.21** | 88.57 | **87.57** | 91.40 | **84.02** | **89.53** |

and CCTV. The CARGO [33] is a large-scale synthetic benchmark with 108,563 images of 5,000 identities from eight ground cameras and five aerial cameras. The LAGPeR [22] contains 63,841 images of 4,231 identities collected by 14 ground cameras and seven aerial cameras. For each benchmark, we strictly follow the official data splits and evaluation protocols for fair comparison. In this context, **G** represents for ground, **A** for aerial, **C** for CCTV, **W** for wearable and **ALL** for view-agnostic protocols. Following previous works, the performance is evaluated by using Cumulative Matching Characteristics (CMC) at Rank-1 and mean Average Precision (mAP) under each benchmark's official testing protocol.

### 4.2 Implementation Details.

Our framework is implemented in PyTorch and trained on an NVIDIA A800 GPU. We adopt the CLIP-ViT-B/16 [20] as the visual and text encoder. For all datasets, images are resized to $256 \times 128 \times 3$. The data augmentation includes random horizontal flipping, random cropping and random erasing. We train for 60 epochs using Adam [9] with batch size 64, with 4 images sampled per identity. The learning rate is set to $3.5 \times 10^{-4}$ for randomly initialized modules and to $5 \times 10^{-6}$ for pre-trained components. A warm-up strategy with 100 iterations is applied before the cosine learning rate schedule. Experimentally, we extract multi-granularity visual tokens from layers $\{5, 8, 12\}$ of the CLIP image encoder, and set $(s_1, s_2) = (0.5, 2.0)$, $c = 1.0$, $\lambda_e = 5$, and $\lambda_g = 1.0/0.25$ for $\mathcal{L}_g/\mathcal{L}_{g,\text{intra}}$ as default. The scaling factors and curvature are optimized end-to-end during training.

### 4.3 Comparison with State-of-the-Arts

As shown in Tab. 1-Tab. 3, we compare our method with state-of-the-arts on the four AG-ReID benchmarks. Our method achieves consistent advantages across all benchmarks. On AG-ReID and AG-ReID v2, it obtains the best mAP under all six evaluation settings. On CARGO, our method also ranks first in mAP across all protocols. Notably, under the challenging A→G setting, it reaches 70.53

**Table 2: Performance comparison on CARGO.** Best results are in **bold**.

| Method | ALL | | A→G | | A→A | | G→G | |
|---|---|---|---|---|---|---|---|---|
| | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| VDT [33] | 55.20 | 64.10 | 42.76 | 48.12 | 66.83 | **82.50** | 71.59 | 82.14 |
| DTST [25] | 55.73 | 64.42 | 43.49 | 50.53 | 63.31 | 80.00 | 72.40 | 78.57 |
| SeCap [22] | 56.89 | 64.72 | 46.37 | 48.75 | 66.90 | **82.50** | 75.24 | 82.54 |
| VIF [7] | 57.46 | 72.76 | 44.55 | 51.25 | 66.98 | **82.50** | 74.19 | 83.93 |
| SD-ReID [26] | 57.47 | 65.06 | 46.44 | 53.12 | 67.70 | **82.50** | 74.08 | 81.25 |
| GSAlign [13] | 57.95 | 65.06 | 61.55 | 64.89 | 65.55 | 80.00 | 73.86 | 83.04 |
| SAS-VPReID [31] | 58.48 | 64.42 | 57.36 | 58.51 | 63.41 | 70.00 | 76.90 | 83.93 |
| LATex [32] | 67.09 | **76.96** | 58.88 | 66.87 | 69.06 | 80.00 | 79.90 | **90.18** |
| **HiHR** | **67.52** | 74.36 | **70.53** | **75.53** | **69.60** | 75.00 | **80.93** | 87.50 |

**Table 3: Performance comparison on LAGPeR.** Best results are in **bold**.

| Method | A→G | | G→A | | G→AG | |
|---|---|---|---|---|---|---|
| | mAP | R1 | mAP | R1 | mAP | R1 |
| Explain [16] | 28.89 | 40.48 | 31.91 | 32.96 | 18.91 | 22.33 |
| VDT [33] | 28.97 | 40.15 | 31.91 | 33.55 | 16.45 | 19.50 |
| SD-ReID [26] | 29.72 | 40.15 | 32.86 | 34.47 | 18.88 | 22.85 |
| SeCap [22] | 30.37 | 41.79 | 33.42 | 35.26 | 19.24 | 24.39 |
| SAS-VPReID [31] | 30.76 | 43.83 | 33.40 | 35.42 | 23.02 | 31.12 |
| **HiHR** | **33.80** | **46.95** | **36.74** | **39.30** | **23.99** | **31.39** |

mAP and 75.53 Rank-1, outperforming the strongest baseline by +8.98 mAP and +8.66 Rank-1. On LAGPeR, our method achieves the best results across all three settings, further validating its effectiveness under diverse aerial-ground view compositions. The most notable advantage appears in mAP, indicating that our method improves the overall ranking quality rather than only the top-1 match. This gain is consistent with our hierarchical alignment strategy, which aggregates view-invariant identity cues while preserving view-specific discriminative details. By avoiding excessive feature collapse under cross-view supervision, our method maintains more reliable similarity ordering with background interference, scale variation, and severe viewpoint changes. Such full-list retrieval improvements suggest strong potential for practical aerial-ground search, where multiple correct candidates beyond the first rank are often valuable.

### 4.4 Ablation Study

In this section, we conduct experiments on CARGO to verify the effect of each design in our framework. We employ the cross-entropy loss and triplet loss to train the baseline model, which is based on the ViT-B/16 from CLIP.

**Component Analysis.** As shown in Tab. 4, adding TMF brings clear performance gains over the baseline, especially under the cross-view A→G setting (+3.44 mAP and +4.26 Rank-1). This improvement verifies the value of text-

**Table 4: Ablation study of each module on CARGO.**

| Method | ALL | | A→G | | A→A | | G→G | |
|---|---|---|---|---|---|---|---|---|
| | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| Baseline | 63.61 | 68.23 | 66.06 | 66.02 | 64.70 | 71.50 | 79.08 | 87.39 |
| Baseline+TMF | 65.46 | 72.08 | 69.50 | 70.28 | 65.64 | 71.50 | 80.22 | **88.29** |
| Baseline+TMF+HHL | **67.52** | **74.36** | **70.53** | **75.53** | **69.60** | **75.00** | **80.93** | 87.50 |

guided multi-granularity fusion in extracting more discriminative features. Further incorporating HHL consistently improves mAP across protocols, showing that hierarchical supervision complements TMF by preserving view-specific discriminative cues while avoiding excessive cross-view collapse. The above results demonstrate the effectiveness of our key modules.

**Hyperparameter Sensitivity Analysis.** As shown in the left panels of Fig. 4, our method maintains stable performance across different settings of the entailment weight $\lambda_e$, curvature initialization $c$, and scaling factors $(s_1, s_2)$. For the entailment weight $\lambda_e$, performance remains consistently stable across a wide range of values. As shown in Fig. 4, mAP fluctuates only slightly around 71–72, and Rank-1 also stays within a narrow performance band. This indicates that our method is insensitive to the precise choice of $\lambda_e$.For curvature $c$, moderate initialization values from 1.0 to 4.0 consistently yield the strong mAP and Rank-1, whereas overly small curvature leads to noticeable degradation. This suggests that sufficient negative curvature is necessary to provide the capacity for hierarchical organization and child-cluster separation. Once such geometry is properly initialized, the model can adapt the curvature during training and is not sensitive to its exact starting value. For the scaling factors $(s_1, s_2)$, different initializations produce similar final accuracy. This indicates that the intended parent-child scale separation can be reliably learned from diverse starting points. Overall, these results show that our method is robust to key hyperparameters and does not require delicate tuning.

**Layer Selection Analysis.** The right panel of Fig. 4 evaluates the effect of selected ViT layers on Multi-Granularity Visual Tokens. Using only the final layer $\{12\}$ is suboptimal, while incorporating intermediate layers improves performance, with $\{5, 8, 12\}$ achieving the best results. This shows that mid-level visual cues provide complementary information to high-level semantic features for AG-ReID. However, the improvement depends on layer complementarity rather than the number of selected layers. The deep-stacked setting $\{10, 11, 12\}$ performs worse than the cross-depth setting $\{5, 8, 12\}$, indicating that adjacent high-level layers contain redundant semantics and lack sufficient granularity diversity. Overly shallow combinations are also less effective, since early layers mainly encode low-level patterns that are weakly aligned with identity-discriminative semantics. Moreover, adding more layers beyond $\{5, 8, 12\}$ does not bring further gains and can even degrade performance, suggesting that excessive fusion introduces redundancy and noisy interactions. These results validate the compact

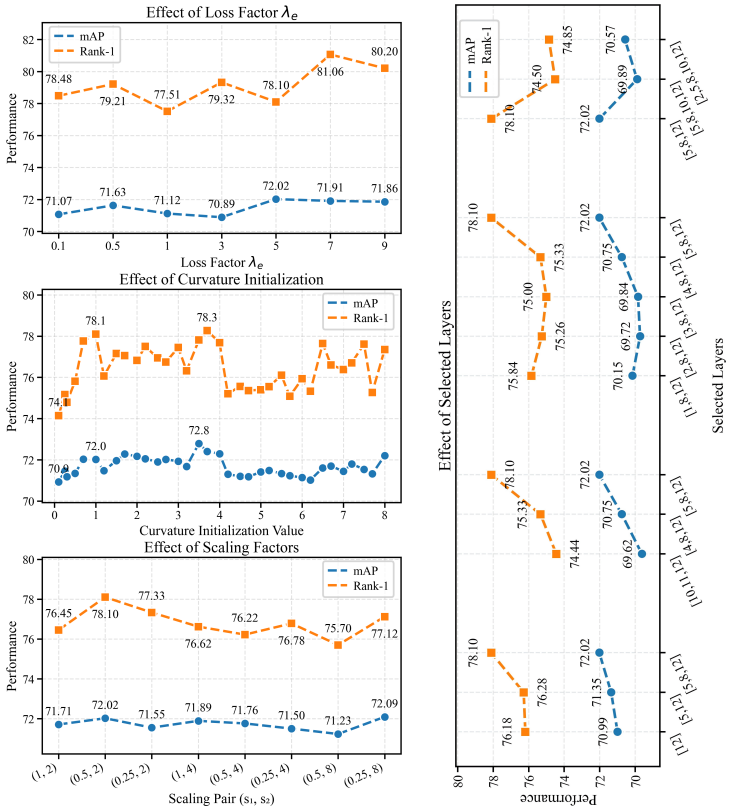


**Fig. 4: Hyperparameter and layer-selection analysis on CARGO.** The left panels report loss factor $\lambda_e$, curvature initialization $c$, and scaling factors $(s_1, s_2)$ from top to bottom. The right panel reports selected ViT layers.

cross-depth design of Multi-Granularity Visual Tokens, which captures complementary semantic and local cues without indiscriminately aggregating all layers.

**Loss Function Analysis.** As shown in Tab. 5, the model without entailment losses achieves the lowest overall performance, indicating that metric supervision alone is insufficient to maintain the hierarchical structure. Introducing $\mathcal{L}_{\mathrm{e}}(\mathbf{z}^{\boldsymbol{\pi}}, \mathbf{z}^{c})$ improves the ALL performance from 64.62/71.15 to 65.46/72.12, showing that constraining child embeddings within parent-induced cones helps reduce child-cluster drift. Adding $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}_P^c)$ further improves the results by enforcing consistency between parent and child prompts. When the vision-language consistency terms $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}^p) + \mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^c, \mathbf{z}^c)$ are also enabled, the model achieves the best performance. These results demonstrate that the four entailment constraints are complementary and jointly necessary for hierarchical representation learning.

**Prompt Design Analysis.** Tab. 6 compares three prompt designs. **Model A** uses fixed handcrafted prompts, where $\mathcal{T}_1$ is "*A photo of a person.*" and $\mathcal{T}_2$ is "*A photo of a person from [an aerial/a ground] view.*". **Model B** adopts CoCoOp [34] to generate instance-conditional prompt tokens, while **Model C** uses

**Table 5: Ablation study of loss functions on CARGO.**

| $\mathcal{L}_{\mathrm{e}}(\mathbf{z}^{\boldsymbol{\pi}}, \mathbf{z}^c)$ | $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}_P^c)$ | $\mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^p, \mathbf{z}^p) + \mathcal{L}_{\mathrm{e}}(\mathbf{z}_P^c, \mathbf{z}^c)$ | ALL | | A→G | | A→A | | G→G | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| | | | 64.62 | 71.15 | 68.06 | 72.34 | 67.42 | **77.50** | 79.80 | 87.50 |
| ✓ | | | 65.46 | 72.12 | 69.50 | 70.21 | 65.64 | 72.50 | 80.22 | **88.39** |
| ✓ | ✓ | | 65.71 | 72.44 | 68.21 | 71.28 | 67.49 | **77.50** | 80.65 | 87.50 |
| ✓ | ✓ | ✓ | **67.52** | **74.36** | **70.53** | **75.53** | **69.60** | 75.00 | **80.93** | 87.50 |

**Table 6: Ablation study of the prompt design on CARGO.**

| Index | Prompt Type | ALL | | A→G | | A→A | | G→G | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| A | Text-only | 64.58 | 68.91 | 67.49 | 71.28 | 66.60 | 75.00 | 80.01 | 85.71 |
| B | Text+CoCoOp | 66.73 | 71.47 | 68.19 | 72.34 | 68.01 | **77.50** | **81.22** | 85.71 |
| C | Text+Coop | **67.52** | **74.36** | **70.53** | **75.53** | **69.60** | 75.00 | 80.93 | **87.50** |

our CoOp [35]-based Dual-Level Prompts. As shown in Tab. 6, Model A obtains reasonable performance, indicating that explicit textual semantics already benefit multi-granularity feature fusion. Model B improves over Model A, showing the effect of learnable prompt adaptation. Compared with Model B, Model C achieves the best overall results, improving ALL performance to 67.52% mAP and 74.36% Rank-1. This suggests that compact dataset-level prompts are more effective for AG-ReID than instance-wise prompt generation. The reason is that aerial-ground discrepancies are mainly governed by systematic view-related variations rather than highly sample-specific changes.

**Fusion Design Analysis.** Tab. 7 compares three fusion designs. **Model A** uses mean aggregation over class tokens from selected layers. **Model B** applies standard cross-attention between prompts and class tokens. **Model C** denotes our proposed TMF design. As shown in Tab. 7, Model A already achieves reasonable performance, confirming the usefulness of multi-layer visual cues. Model B improves over Model A, indicating that adaptive prompt-token interaction is more effective than static averaging. Our TMF obtains the best results, outperforming Model B by +1.72 mAP and +3.21 Rank-1 under the ALL setting. This verifies that the proposed text-guided multi-granularity fusion more effectively exploits complementary information across layers for AG-ReID.

## 5 Conclusion

In this paper, we address AG-ReID under severe aerial-ground viewpoint discrepancies, where overly strong cross-view alignment may suppress view-specific yet discriminative cues. To this end, we propose HiHR, a hierarchical hyperbolic representation framework that jointly exploits view-invariant identity semantics and view-specific discriminative details. With Dual-Level Prompts and Multi-Granularity Visual Tokens as semantic and visual bases, TMF performs text-guided multi-granularity fusion and fine-grained injection to produce complementary feature embeddings. HHL further organizes these embeddings in a

**Table 7: Ablation study of the fusion design on CARGO.**

| Index | Fusion Type | ALL | | A→G | | A→A | | G→G | |
|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | mAP | R1 | mAP | R1 | mAP | R1 |
| A | Mean | 64.95 | 69.87 | 66.86 | 69.15 | 65.49 | 72.50 | 80.16 | 87.50 |
| B | CrossAtt | 65.80 | 71.15 | 67.08 | 70.21 | 67.20 | **77.50** | 80.75 | 86.61 |
| C | TMF | **67.52** | **74.36** | **70.53** | **75.53** | **69.60** | 75.00 | **80.93** | **87.50** |

hyperbolic hierarchy, where the parent level promotes identity separability and cross-view consistency, while the child level preserves view-conditioned refinement. Extensive experiments on four AG-ReID benchmarks validate the effectiveness of the proposed components and demonstrate competitive or state-of-the-art performance, reflecting improved full-ranking retrieval quality.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62576069) and Natural Science Foundation of Liaoning Province (No.2025-MS-025). The author sincerely thanks Prof. Pingping Zhang for his valuable guidance and continuous support throughout this work.

## References

1. Cannon, J.W., Floyd, W.J., Kenyon, R., Parry, W.R., et al.: Hyperbolic geometry. Flavors of geometry **31**(59-115), 2 (1997)
2. Fang, P., Harandi, M., Petersson, L.: Kernel methods in hyperbolic spaces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10665–10674 (2021)
3. Ha, R., Jiang, S., Li, B., Pan, B., Zhu, Y., Zhang, J., Zhu, X., Gong, S., Wang, J.: Multi-modal multi-platform person re-identification: Benchmark and method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10251–10261 (2025)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification
5. Huang, Y., Chen, H., Feng, Z., Lai, J.: Perspective driven prototype alignment for aerial-ground person re-identification. In: International Conference on Image and Graphics. pp. 517–528. Springer (2025)
6. Huang, Z., Liu, X.: Generalizable object re-identification via visual in-context prompting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22539–22550 (2025)
7. 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)
8. Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., Lempitsky, V.: Hyperbolic image embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6418–6428 (2020)

9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
10. Lee, J.M.: Riemannian manifolds: an introduction to curvature. Springer Science & Business Media (2006)
11. Li, H., Wang, Y., Hao, W., Zhang, P., Wang, D., Lu, H.: Ragtrack: Language-aware rgbt tracking with retrieval-augmented generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28179–28189 (2026)
12. Li, H., Wang, Y., Hu, X., Hao, W., Zhang, P., Wang, D., Lu, H.: Cadtrack: Learning contextual aggregation with deformable alignment for robust rgbt tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6109–6117 (2026)
13. Li, Q., Li, J., Zhang, Y., Tan, L., Chen, J., Ji, J.: Gsalign: Geometric and semantic alignment network for aerial-ground person re-identification. Advances in Neural Information Processing Systems **38**, 93785–93805 (2026)
14. Li, S., Sun, L., Li, Q.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1405–1413 (2023)
15. Mao, D., Teng, S., Lyu, X.: Cvaf: A clip-based view-consistent alignment framework for aerial-ground person re-identification. ACM Transactions on Multimedia Computing, Communications and Applications **22**(3), 1–19 (2026)
16. Nguyen, H., Nguyen, K., Sridharan, S., Fookes, C.: Aerial-ground person re-id. In: 2023 IEEE International Conference on Multimedia and Expo. pp. 2585–2590 (2023)
17. 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)
18. 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: IEEE International Joint Conference on Biometrics. pp. 1–10 (2023)
19. Pal, A., Van Spengler, M., D'Amely di Melendugno, G., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. In: International Conference on Learning Representations. vol. 2025, pp. 87371–87399 (2025)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
22. 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)
23. Wang, Y., Zhang, P., Sun, C., Wang, D., Lu, H.: What makes you unique? attribute prompt composition for object re-identification. IEEE Transactions on Circuits and Systems for Video Technology pp. 3173–3184 (2025)
24. Wang, Y., Zhang, P., Wang, D., Lu, H.: Other tokens matter: Exploring global and local features of vision transformers for object re-identification. Computer Vision and Image Understanding **244**, 104030 (2024)

25. Wang, Y., Pishgar, M.: Dynamic token selection for aerial-ground person re-identification. arXiv preprint arXiv:2412.00433 (2024)
26. 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)
27. Wang, Y., Liu, Y., Zheng, A., Zhang, P.: Decoupled feature-based mixture of experts for multi-modal object re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8141–8149 (2025)
28. Wang, Y., Lv, Y., Zhang, P., Lu, H.: Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29701–29710 (2025)
29. Wang, Y., Zhang, P., Liu, X., Tu, Z., Lu, H.: Unity is strength: Unifying convolutional and transformeral features for better person re-identification. IEEE Transactions on Intelligent Transportation Systems pp. 3713–3723 (2025)
30. Yan, P., Liu, X., Zhang, P., Lu, H.: Learning convolutional multi-level transformers for image-based person re-identification. Visual Intelligence **1**(1), 24 (2023)
31. Yang, Q., Zhang, P., Wang, Y., Gong, Z.: Sas-vpreid: A scale-adaptive framework with shape priors for video-based person re-identification at extreme far distances. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 1599–1608 (March 2026)
32. Zhang, P., Hu, X., Wang, Y., Lu, H.: Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. arXiv preprint arXiv:2503.23722 (2025)
33. 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)
34. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
35. 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)
36. Zhu, R., Chen, H., Xie, X., Lai, J.: Global-local prompts-driven semantic guidance for aerial-ground person re-identification. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 99–112. Springer (2025)