Text-based Aerial-Ground Person Retrieval

Xinyu Zhou [1], Yu Wu [1], Jiayao Ma [2], Wenhao Wang [2], Min Cao* [1], Mang Ye [3]

[1]School of Computer Science and Technology, Soochow University
[2]AgiBot
[3]School of Computer Science, Wuhan University
xyzhou2023@stu.suda.edu.cn, caomin0719@126.com

Abstract

This work introduces Text-based Aerial-Ground Person Retrieval (TAG-PR), which aims to retrieve person images from heterogeneous aerial and ground views with textual descriptions. Unlike traditional Text-based Person Retrieval (T-PR), which focuses solely on ground-view images, TAG-PR introduces greater practical significance and presents unique challenges due to the large viewpoint discrepancy across images. To support this task, we contribute: (1) TAG-PEDES dataset, constructed from public benchmarks with automatically generated textual descriptions, enhanced by a diversified text generation paradigm to ensure robustness under view heterogeneity; and (2) TAG-CLIP, a novel retrieval framework that addresses view heterogeneity through a hierarchically-routed mixture of experts module to learn view-specific and view-agnostic features and a viewpoint decoupling strategy to decouple view-specific features for better cross-modal alignment. We evaluate the effectiveness of TAG-CLIP on both the proposed TAG-PEDES and existing T-PR benchmarks.

Code — https://github.com/Flame-Chasers/TAG-PR

Introduction

Text-based Person Retrieval (T-PR) (Li et al. 2017b) is a specialized vision-language learning task that identifies person images using natural language descriptions. It can be viewed as an extension of the traditional person re-identification (Re-ID) (Wang et al. 2022; Liu, Ye, and Du 2024), which matches person images across different cameras, by replacing the query image with text. T-PR has gained increasing academic attention in recent years (Bai et al. 2023b; Song, Hu, and Zhao 2024; Tan et al. 2024; Bai et al. 2025) due to its potential in applications such as suspect tracking and surveillance. Current research focuses on achieving effective cross-modal alignment between person images and texts, and has made substantial breakthroughs in performance (Qin et al. 2024; Jiang et al. 2025; Cao et al. 2025).

Nevertheless, these works focus mainly on scenarios involving only ground-view camera networks, assuming that all person images are captured from low-altitude, ground-view cameras. This assumption does not fully reflect real-world conditions, where advances in airborne platforms and imaging technologies have enabled widespread deployment of aerial cameras alongside traditional ground systems. High-altitude aerial-view cameras offer broader coverage and complementary perspectives to ground-based cameras, and the synergy between these two viewpoints becomes increasingly important in real-world retrieval scenarios. In response, the Re-ID community has explored in mixed ground-aerial settings (Zhang et al. 2023; Nguyen et al. 2024; Wang et al. 2025a), such efforts remain absent in T-PR task. For this, we focus on a more practical setting in this paper, Text-based Aerial-Ground Person Retrieval (TAG-PR), which involves retrieving person images captured from heterogeneous aerial and ground viewpoints

Figure 1: Illustration of TAG-PR. It aims to retrieve a target individual from an image gallery containing heterogeneous-view images, given a query text. The gallery includes ground-view images, typically captured by CCTV cameras at low altitudes (below 10 meters), and aerial-view images taken by UAVs from significantly higher altitudes ($\gg 10m$).

*Corresponding Author

based on textual descriptions, as shown in Figure 1.

Existing T-PR datasets primarily consist of ground-view person images, lacking datasets specifically tailored for TAG-PR. To address this gap, we firstly introduce TAG-PEDES, a dataset explicitly designed for TAG-PR. It is economically constructed by collecting person images from ground and aerial views based on public Re-ID datasets (Zheng et al. 2015; Zhang et al. 2020; Li et al. 2021b; Nguyen et al. 2024; Zhang et al. 2023) with automatically generated textual descriptions. A central challenge lies in generating high-quality descriptions for these images that exhibit significant heterogeneity due to drastic view-angle differences. For this, we propose a Diversified Text Generation (DTG) paradigm powered by multimodal large language models. DTG integrates multiple distinct text generation strategies to ensure robustness and adaptability of high-quality textual descriptions across view-angle images.

In addition, we propose a novel method tailored to this task. Traditional T-PR methods focus on cross-modal alignment between image and text. While in more complex TAG-PR, it is equally important to account for view heterogeneity inherent in the image modality. Ground-view images, captured at eye level, present frontal or side views of a person, whereas aerial-view images, taken from high altitudes, offer a top-down perspective, leading to distinct visual characteristics. For this, we propose TAG-CLIP. It explicitly incorporates view heterogeneity during image encoding and cross-modal alignment. 1) For image encoding, the images from different viewpoints exhibit distinct visual characteristics while also sharing view-agnostic features. A natural way is to extract view-specific patterns separately while jointly learning view-agnostic features. For this, we design a Hierarchically-Routed Mixture of Experts (HR-MoE) module and integrate it into the image encoder. It comprises a two-tiered routing mechanism, along with two specialized expert groups, all of which are designed to process view-specific and view-agnostic features separately. 2) For cross-modal alignment learning, unlike images, textual descriptions lack view-specific information[1], making it crucial to decouple view-specific features from the image feature before alignment. We propose a viewpoint decoupling strategy that eliminates viewpoint-specific cues from image features. The resulting view-agnostic visual features are then aligned with textual features for better performance.

In summary, our contributions are as follows. 1) We introduce text-based aerial-ground person retrieval, a cross-modal and cross-platform task with broader application potential than traditional text-based person retrieval. 2) We construct a dataset, TAG-PEDES, for this task, supported by the proposed diversified text generation paradigm to collect high-quality textual descriptions from varied viewpoints. 3) We propose TAG-CLIP, a novel method which incorporates a hierarchically-routed mixture of experts in the image encoder and a viewpoint decoupling strategy to better learn and align view-specific and view-invariant features.

[1]TAG-PR retrieves target individuals based on textual descriptions that typically focus on appearance-related attributes and do not include viewpoint cues in real-world inference scenarios.

Related Work

Text-based Person Retrieval

T-PR has emerged as an active research area in recent years, marked by notable methodological advances. Early approaches (Li et al. 2017b; Wang et al. 2017; Chen et al. 2021; Li et al. 2017a) employ independently trained unimodal encoders (e.g., VGG (Simonyan and Zisserman 2014) for images and LSTM (Graves and Graves 2012) for text) to extract global features for cross-modal alignment, but struggle to capture fine-grained semantic correspondences. Subsequent works (Jing et al. 2020; Fujii and Tarashima 2023; Wang et al. 2020) address this limitation by incorporating explicit alignment mechanisms, often relying on external tools to align fine-grained entities. More recently, vision-language pretraining models (Radford et al. 2021; Li et al. 2021a) have attracted significant attention due to their strong cross-modal representation capabilities. Leveraging these models, a series of studies (Jiang and Ye 2023; Cao et al. 2024; Yan et al. 2024; Yang et al. 2023) have achieved remarkable performance gains. For example, Cao *et al.* (Cao et al. 2024) conducted comprehensive evaluations of CLIP's fine-grained alignment capacity. However, these works focus mainly on text-based retrieval for ground-view images, limiting their real-world applicability. In contrast, this work introduces TAG-PR. A concurrent work AEA-FIRM (Wang et al. 2025b) also studies this task but differs: 1) Task Definition. AEA-FIRM defines on strictly paired ground-aerial images per identity, whereas we allow diverse combinations (ground-only, aerial-only, and mixed-view), better reflecting real-world scenarios. 2) Dataset Construction. AEA-FIRM generates textual descriptions solely for ground views and reuses them for aerial views, ignoring viewpoint-specific cues. We propose DTG to enable viewpoint-aware descriptions. 3) Modeling View Heterogeneity. AEA-FIRM employs an AEA loss with a rigid triplet input structure, offering limited flexibility to view differences. We propose TAG-CLIP, which explicitly models view heterogeneity during both image encoding and cross-modal alignment, leading to better performance.

Aerial-Ground Person Re-ID

Aerial-ground person Re-ID has recently emerged as a challenging task. Nguyen *et al.* (Nguyen et al. 2023) introduced the first benchmark, AG-ReID, and proposed a two-stream framework combining a transformer-based Re-ID stream with an explainable stream, using attribute supervision to improve feature discriminability. They later extended the work with the larger AG-ReID.v2 dataset and a more robust three-stream network for improved performance (Nguyen et al. 2024). However, these methods rely on one-hot attribute labels, which may limit generalizability due to their dependence on manual semantic cues. In parallel, Zhang *et al.* (Zhang et al. 2024) introduced a view-decoupled transformer, yet neglects fine-grained local viewpoint disentanglement crucial for cross-modal alignment. In contrast, our proposed HR-MoE distinguishes view-specific and view-agnostic local features by assigning them to different expert groups, enabling more robust feature extraction.

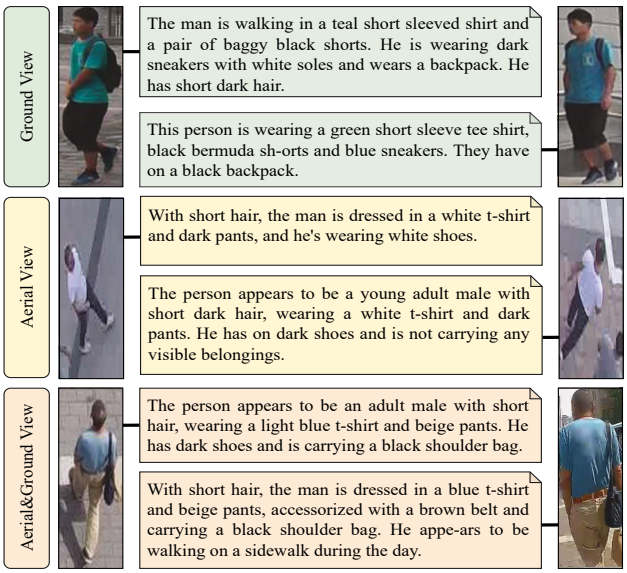

Figure 2: Examples from the proposed TAG-PEDES dataset. More examples are provided in the Appendix.

Dataset

Dataset Construction

To support the TAG-PR task, we construct a benchmark dataset TAG-PEDES, by collecting person images from existing Re-ID datasets and generating corresponding textual descriptions. In addition, to ensure test data quality, we manually review and refine the textual annotations of the test set.

Image Collection. To better reflect the complex viewpoint diversity present in real-world scenarios, we curate a collection from multiple Re-ID datasets. Specifically, we obtain ground-view images from ground-only dataset (Zheng et al. 2015), aerial-view images from aerial-only datasets (Zhang et al. 2020; Li et al. 2021b), and multi-view images from aerial-ground datasets (Nguyen et al. 2024; Zhang et al. 2023), with view labels (*i.e.*, aerial or ground) provided in the original datasets. After filtering out low-resolution images ($\leq 1,000$ pixels), we obtain an image gallery containing $28,206$ images of $6,840$ unique identities captured from multiple viewpoints and cameras. Table 1 presents the detailed composition of the image gallery of TAG-PEDES. As shown in Figure 2, the diversity in viewpoints introduces variations in image resolution and pedestrian postures, making the dataset more representative of real-world scenarios, while posing challenges for generating high-quality texts.

Text Generation. To address the challenge posed by various image viewpoints, we propose a Diversified Text Generation (DTG) paradigm powered by Multimodal Large Language Models (MLLM) (Li et al. 2024a) to enable automated image annotation. Specifically, DTG comprises three distinct text generation strategies.

Prompt-based generation is the most straightforward strategy. Specifically, we activate the powerful text-generation capabilities of the MLLM (Wang et al. 2024) with a fixed, empirically designed prompt:

Don't mention the background of the people in the image. Please provide a detailed description of this person's age, gender, top (including color and style), bottom (including color and style), hair (including color and style), shoes (including color and style), belongings (including color and style). Finally, combine all the details into a single sentence.

Template-based generation provides finer control over prompt-based generation by leveraging auxiliary templates to direct the model's attention to key pedestrian attributes. we provide the MLLM with the prompt (Tan et al. 2024):

Generate a description about the overall appearance of the person, in a style similar to the template: {template}. If some requirements in the template are not visible, you can ignore. Do not imagine any contents not in the image.

The placeholder *{template}* is replaced with specific templates presented in the Appendix.

Attribute-based generation provides the MLLM with manually annotated attribute labels (*e.g.*, young, male) from Re-ID datasets, which offer explicit descriptions of pedestrian's age, gender, clothing and so on. These attributes are integrated into the prompt of the template-based generation to improve the accuracy and detail of textual descriptions:

Generate a description about the overall appearance of the person, based on the attribute: [label], in a style similar to the template: [template]. If some requirements in the template are not visible, you can ignore. Do not imagine any contents that are not in the image.

Using the three aforementioned text generation strategies, we generate two textual descriptions per image. For images with attribute annotations (*e.g.,* AG-ReID.v2 (Nguyen et al. 2024), UAVHuman (Li et al. 2021b)), we combine attribute-based generation with either prompt-based or template-based generation, selected randomly. For images without attribute labels, we use a mix of prompt- and template-based generation strategies. This design ensures both the quality and diversity of textual descriptions.

Test Set Correction. Given the inevitable noise in generated texts, we ensure the reliability of evaluation by employing six qualified annotators to manually inspect and revise the generated descriptions in the test set. This process enhances the accuracy and quality of the test data.

View	Image Source	#IDs	#Images
Ground	Market-1501 (Zheng et al. 2015)	1,100	3,300
Aerial	PRAI-1581 (Zhang et al. 2020)	844	3,374
	UavHuamn (Li et al. 2021b)	655	2,476
Aerial&Ground	AG-Reid.v2 (Nguyen et al. 2024)	1,615	7,750
	G2APS (Zhang et al. 2023)	2,626	11,306

Table 1: Composition of the image gallery in TAG-PEDES. All view labels are inherited from the original datasets.

Dataset Analysis

We first conduct a quality assessment on TAG-PEDES to evaluate its reliability. To measure the alignment between generated descriptions and paired images, we define four

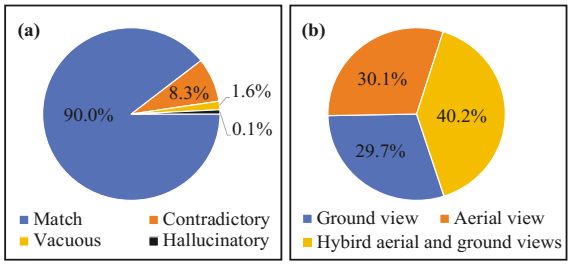

Figure 3: (a) Results of quality assessment on TAG-PEDES. (b) The proportion of identities from different viewpoints.

evaluation categories: *Match*, *Contradictory*, *Hallucinatory*, and *Vacuous*. We use a crafted instruction to guide an MLLM (Wang et al. 2024) for classification. As shown in Figure 3 (a), approximately 90% of the descriptions are classified as *Match*, indicating strong image-text alignment. Further details are provided in the Appendix.

We next highlight several distinctive characteristics of the proposed TAG-PEDES dataset. 1) **Viewpoint Diversity.** Unlike traditional T-PR datasets that primarily contain ground-view images, TAG-PEDES includes a balanced distribution of identities across diverse viewpoints, as shown in Figure 3 (b). 2) **High-Quality Synthetic Descriptions.** TAG-PEDES provides fine-grained textual descriptions averaging 39 words in length, generated by the proposed DTG and manually refined for test set. Compared to manual-only annotations in prior datasets, this ensures scalability. 3) **Increased Challenge.** The dataset poses difficulty by requiring to address both cross-modal alignment and multi-view heterogeneity, making it more challenging than existing Re-ID and T-PR benchmarks. A comprehensive comparison with existing Re-ID and T-PR datasets is shown in the Appendix.

Method

In this section, we elaborate on the proposed TAG-CLIP, as shown in Figure 4. We begin with a brief review of the baseline model, followed by a detailed introduction of the proposed HR-MoE module and viewpoint decoupling strategy.

Baseline

We adopt TBPS-CLIP (Cao et al. 2024) as our baseline, a lightweight yet effective method for T-PR. It consists of an image encoder and a text encoder to extract visual and textual features, respectively. The image encoder, consisting of 12 Vision Transformer (ViT) blocks, takes an image I and divides it into S_I patches, which are linearly projected and combined with positional embeddings. A [CLS] token and a view token are prepended to the patch sequence before being fed into the encoder. The view token supports the HR-MoE module and viewpoint decoupling strategy detailed in the next section. This yields image feature $F_v = \{v_{cls}, v_1, \ldots, v_{S_I}, v_{view}\}$, where v_{cls} serves as the global visual feature, v_i $(i = 1, \cdots, S_I)$ are local features, and v_{view} is the view feature. Similarly, the input text is tokenized and augmented with [SOS] and [EOS] tokens to mark the start and end of text, and then passed through a Transformer-based text encoder to obtain textual features $F_t = \{t_{sos}, t_1, \ldots, t_{S_T}, t_{eos}\}$. t_{eos} serves as the global text feature and t_i $(i = 1, \cdots, S_T)$ are the local textual features.

The global alignment is then applied on global features:

$$L_{GA} = L_{N-ITC}(v_{cls}, t_{eos}) + L_{R-ITC}(v_{cls}, t_{eos}). \quad (1)$$

The loss $L_{N-ITC}(\cdot)$ represents the alignment of image-text pairs by encouraging high similarity for positive pairs, while $L_{R-ITC}(\cdot)$ focuses on pushing apart negative pairs by aligning the similarity with the label distribution in reverse. Further details are provided in (Cao et al. 2024). In addition, we incorporate a Token Selection Embedding (TSE) module (Qin et al. 2024), which selects and aggregates key local features, producing t_{tse} for text and v_{tse} for vision. Then, local alignment which is performed as follows:

$$L_{LA} = L_{N-ITC}(v_{tse}, t_{tse}) + L_{R-ITC}(v_{tse}, t_{tse}). \quad (2)$$

Furthermore, we introduce an ID loss (Zheng et al. 2020) L_{id}, which enforces samples sharing the same identity to be clustered into a single class, promoting intra-class compactness and inter-class separability in the learned feature space.

Hierarchically-Routed Mixture of Experts

To learn robust visual features from view-heterogeneous images, we introduce a Hierarchically-Routed Mixture of Experts (HR-MoE) module into a subset of ViT blocks in the image encoder, termed HR-MoE Blocks (see Figure 4).

The HR-MoE Block processes image features F_v with a self-attention layer followed by the HR-MoE module. The module consists of two hierarchical routers and a set of experts $\{\mathcal{E}_i \mid i \in [1, N_e]\}$, which are divided into two groups specialized for aerial and ground views. The hierarchical routing mechanism allows each image feature to be dynamically routed to the most appropriate experts, enabling progressive learning of both view-specific and view-agnostic features from images captured under diverse viewpoints.

Specifically, the first router, referred to as the image level router $\mathcal{R}_{Img}$, predicts the image's view category,

$$g_{img} = \mathcal{R}_{Img}(v_{view}), \quad (3)$$

$$z = \arg\max\left(Softmax(g_{img})\right), \quad (4)$$

where $\mathcal{R}_{Img}$ is implemented as a linear classifier, g_{img} is the output one-hot logits, and z is the predicted binary view label (0 for aerial, 1 for ground) and will be used to route the image features to the corresponding expert group in Eq. 7 for subsequent specialized feature processing.

The second routing layer, termed the feature-level router $\mathcal{R}_{Feat}$, performs a finer-grained assignment for each visual feature by selecting the most appropriate experts. Formally, for the i-th visual feature F_v^i in the image features F_v, the routing process computes a probability distribution $P \in \mathbb{R}^{N_e}$ over all experts:

$$g_{feat} = \mathcal{R}_{Feat}(F_v^i), \quad (5)$$

$$P = Softmax(g_{feat} + M). \quad (6)$$

Here, $\mathcal{R}_{Feat}$ is implemented as a linear classification layer that outputs expert assignment logits $g_{feat} \in \mathbb{R}^{N_e}$, where

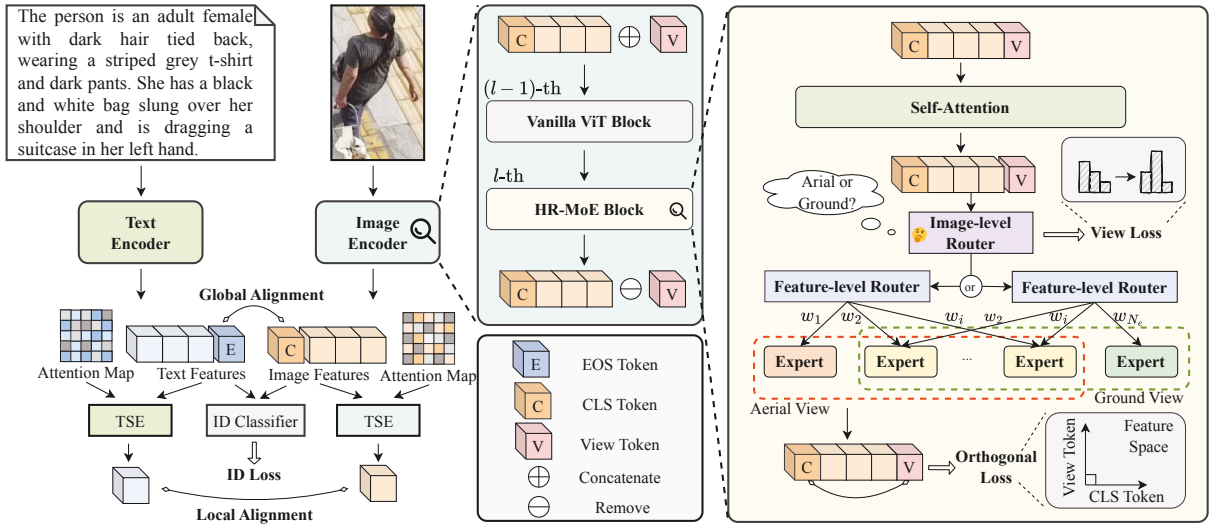

Figure 4: Overview of TAG-CLIP. It comprises an image encoder and a text encoder. Several ViT blocks in the image encoder are augmented with the HR-MoE module, which employs hierarchical routers and expert groups for robust visual feature extraction. A viewpoint decoupling strategy, consisting of two loss functions, is used to decouple viewpoint information from global visual features, thereby improving cross-modal alignment.

each element indicates the compatibility between feature F_v^i and each expert. $M \in \mathbb{R}^{N_e}$, as a binary mask, is used to restrict expert selection based on the predicted view label z, obtained from the image-level router $\mathcal{R}_{Img}$.

Prior to defining M, we detail the expert group partitioning. The N_e experts is partitioned into two groups $E_{aerial} = \{\mathcal{E}_n \mid n \in [1, e_0]\}$ for aerial-view images and $E_{ground} = \{\mathcal{E}_n \mid n \in [e_1, N_e]\}$ for ground-view images. Notably, $e_1 < e_0$, which implies a partial overlap between the two expert groups. The shared experts $\{\mathcal{E}_n \mid n \in [e_1, e_0]\}$ are designed to capture view-agnostic features, while the non-overlapping ones in each group specialize in view-specific features. Hereby, we define j-th element of the mask M,

$$M_j = \begin{cases} 0, & \text{if } z = 0 \text{ and } j \in [1, e_0], \\ 0, & \text{if } z = 1 \text{ and } j \in [e_1, N_e], \\ -\infty, & \text{otherwise.} \end{cases} \tag{7}$$

This masking strategy enforces view-specific expert selection by setting certain entries in g_{feat} to $-\infty$, forcing their softmax outputs to zero and disabling the corresponding experts. In contrast, entries set to 0 remain unchanged, allowing the router to choose experts based on learned preferences within the permitted group.

Next, we perform a Top-K selection over the expert set for the i-th feature to identify the K most relevant experts according to the computed probability distribution P. Let $\{w_1, \ldots, w_K\}$ denote their corresponding normalized routing weights. The final updated features F_v^i of the i-th feature is obtained via a weighted summation of the expert outputs:

$$F_v^i = \sum_{k=1}^{K} w_k \cdot \mathcal{E}_k(F_v^i), \tag{8}$$

where $\mathcal{E}_k(\cdot)$ denotes the k-th expert. The resulting vector F_v^i serves as the refined feature and is forwarded to the next layer for further processing.

Viewpoint Decoupling Strategy

To perform effective cross-modal alignment, it is critical to address the inherent asymmetry between visual and textual features, that is, image features contain explicit viewpoint information, whereas textual descriptions typically lack such cues. In response, we propose a viewpoint decoupling strategy, comprising a view loss and an orthogonal loss.

The view loss trains the view feature v_{view} to explicitly capture viewpoint information from images. Formally, we minimize the cross-entropy between the logits g_{img} in Eq. 3 and the one-hot ground-truth view label vector z_{gt}:

$$L_{view} = \mathcal{H}(g_{img}, z_{gt}), \tag{9}$$

where $\mathcal{H}$ represents the cross-entropy computation.

The orthogonal loss is used to decouple viewpoint-specific information from the global feature v_{cls} by enforcing orthogonality between the view feature v_{view} and the global feature v_{cls} in feature space. The loss is defined as:

$$L_{ortho} = \min\left(|\cos(v_{cls}, v_{view})|, \alpha\right), \tag{10}$$

where $|\cdot|$ denotes the absolute value, cos represents cosine similarity, and α is a predefined threshold. This geometric constraint encourages orthogonality between the feature vectors to effectively decouple most viewpoint-specific information from v_{cls}. The threshold α, in turn, prevents the exclusion of viewpoint-specific yet highly discriminative pedestrian features from v_{cls}. As a result, the global visual feature v_{cls} retains both viewpoint agnosticism and discriminability, making it well-suited for cross-modal alignment.

Training and Inference

In the training phase, we jointly optimize both alignment-oriented and viewpoint-decoupling loss functions:

$$L = L_1 + L_2, \tag{11}$$

$$L_1 = L_{GA} + L_{LA} + \lambda_{id} L_{id}, \tag{12}$$

$$L_2 = L_{view} + \lambda_{ortho} L_{ortho}. \tag{13}$$

Here, λ_{id} and λ_{ortho} serve as hyperparameters.

In inference, the cross-modal similarity score sim is computed as the average of global and local alignment scores:

$$sim = \frac{1}{2}[\cos(v_{cls}, t_{eos}) + \cos(v_{tse}, t_{tse})]. \tag{14}$$

Experiments

We evaluate our method on four datasets: our newly proposed TAG-PEDES, along with three widely used T-PR datasets CUHK-PEDES (Li et al. 2017a), ICFG-PEDES (Ding et al. 2021) and RSTPReID (Zhu et al. 2021).

TAG-PEDES has person images captured from diverse viewpoints, comprising $28,206$ images of $6,840$ identities, each paired with 2 textual descriptions. The dataset is divided into a training set and a test set. The training set consists of $19,954$ images from $4,840$ identities and $39,908$ text descriptions. The test set includes $8,252$ images, $16,504$ text descriptions, and $2,000$ identities.

Details on the T-PR datasets, evaluation metrics, and implementation are provided in the Appendix.

Comparison with State-of-the-Art Methods

We evaluate the proposed method on the newly constructed TAG-PEDES dataset, aiming to evaluate its effectiveness. As shown in Table 2, we compare TAG-CLIP with several representative T-PR methods, as well as the concurrent TAG-PR method AEA-FIRM. All competing methods are reproduced on TAG-PEDES dataset using their publicly available code. Firstly, our method outperforms all T-PR methods, regardless of whether they are based on CLIP, achieving the best performance on both R@1 and mAP metrics. Specifically, TAG-CLIP surpasses the state-of-the-art method RDE (Qin et al. 2024) by 1.06% and 1.24% in R@1 and mAP, respectively. Secondly, compared to AEA-FIRM, which is specialized for TAG-PR, our method demonstrates superior performance. AEA-FIRM employs an AEA loss with a strict triplet input structure, which limits its adaptability to view variations, whereas our TAG-CLIP explicitly models view heterogeneity during both image encoding and cross-modal alignment, resulting in better performance.

In addition, we evaluate TAG-CLIP on the conventional ground-only dataset CUHK-PEDES and compare it with state-of-the-art methods, as shown in Table 3. **It is reasonable that TAG-CLIP exhibits slightly inferior performance in this setting, due to the following factors.** First, under the single-view setting where all images are captured from the ground viewpoint, both the HR-MoE module and the viewpoint decoupling strategy in TAG-CLIP become inapplicable. Consequently, TAG-CLIP essentially degenerates into a baseline model, resulting in suboptimal performance compared to methods ***specialized for single-view scenarios***. Second, our dual-stream CLIP-based architecture lacks the advantage of cross-modal fusion modules leveraged by one-stream models such as AUL (Li et al. 2024b), which further contributes to the performance gap.

Methods	R@1	R@5	R@10	mAP
w/o CLIP:				
RaSa (Bai et al. 2023a)	61.32	78.31	84.03	48.77
APTM (Yang et al. 2023)	61.52	78.25	83.94	47.77
AUL (Li et al. 2024b)	59.26	76.64	82.89	45.79
w/ CLIP:				
CFine (Yan et al. 2023)	56.68	74.78	81.34	-
IRRA (Jiang and Ye 2023)	59.61	77.52	83.56	46.53
TBPS-CLIP (Cao et al. 2024)	60.28	77.71	83.70	46.45
RDE (Qin et al. 2024)	61.58	78.54	84.08	48.03
AEA-FIRM (Wang et al. 2025b)	51.83	71.78	79.01	38.60
TAG-CLIP (Ours)	**62.64**	**79.10**	**84.81**	**49.27**

Table 2: Comparison with state-of-the-art methods on TAG-PEDES. We reproduce these compared methods with their official released code.

Methods	R@1	R@5	R@10	mAP
w/o CLIP:				
LGUR (Shao et al. 2022)	65.25	83.12	89.00	-
RaSa (Bai et al. 2023a)	76.51	90.29	94.25	**69.38**
APTM (Yang et al. 2023)	76.53	90.04	94.15	66.91
AUL (Li et al. 2024b)	**77.23**	**90.43**	**94.41**	-
w/ CLIP:				
CFine (Yan et al. 2023)	69.57	85.93	91.15	-
IRRA (Jiang and Ye 2023)	73.38	89.93	93.71	66.13
Propot (Yan et al. 2024)	74.89	89.90	94.17	67.12
TBPS-CLIP (Cao et al. 2024)	72.66	88.14	92.72	64.97
RDE (Qin et al. 2024)	75.94	90.14	94.12	67.56
TAG-CLIP (Ours)	74.38	88.30	92.59	67.18

Table 3: Comparison with state-of-the-art methods on CUHK-PEDES.

In the Appendix, we present evaluations of TAG-CLIP on ground-only datasets (ICFG-PEDES and RSTPReid) and aerial-only setting to further demonstrate its effectiveness.

Ablation Study

In this section, we perform ablation studies to evaluate the effectiveness of the proposed HR-MoE module and viewpoint decoupling strategy in TAG-CLIP.

Ablation on HR-MoE. We propose HR-MoE, integrated into the image encoder to extract robust visual features from heterogeneous image views. To verify its effectiveness, we compare the HR-MoE block with two alternative configurations: the standard ViT block and the vanilla MoE structure. The former refers to the baseline where HR-MoE is not applied to the image encoder, while the latter incorporates a set of experts for feature learning without employing the hierarchical routing mechanism. The results are summarized in Table 4. First, when comparing the ViT block with HR-MoE (No.1 *vs.* No.3), we observe performance gains, demonstrat-

No.	Block			Decoupling		R@1	R@5	R@10	mAP
	ViT	MoE	HR-MoE	L_{view}	L_{ortho}				
1	✓					59.09	76.58	82.55	48.05
2		✓				60.26	78.01	84.06	48.95
3			✓			61.60	78.45	84.39	48.94
4			✓	✓		61.55	78.41	84.08	48.73
5			✓		✓	62.32	**79.37**	84.51	49.22
6			✓	✓	✓	**62.64**	79.10	**84.81**	**49.27**

Table 4: Ablation results of TAG-CLIP's components.

No.	Train Set	Test Set	R@1	R@5	R@10	mAP
1	Aerial	Aerial	54.04	73.74	80.61	51.45
2	Aerial&Ground	Aerial	54.64	73.01	80.40	51.77
3	Ground	Ground	63.57	80.93	86.29	60.31
4	Aerial&Ground	Ground	64.49	80.57	86.55	60.76
5	Aerial	Aerial&Ground	47.96	69.48	76.76	36.02
6	Ground	Aerial&Ground	53.48	73.03	79.44	41.19
7	Aerial&Ground	Aerial&Ground	57.80	76.90	82.58	44.40

Table 5: Ablation results of diverse viewpoints.

ing the effectiveness of the HR-MoE design. Second, in contrast to the vanilla MoE (No.2 *vs.* No.3), HR-MoE achieves a more substantial improvement. This result suggests that our hierarchical routing strategy and expert grouping mechanism are critical to achieving superior performance. We further visualize features from ViT, MoE, and HR-MoE with t-SNE (Fig. 5), showing that HR-MoE yields more discriminative features with reduced viewpoint sensitivity.

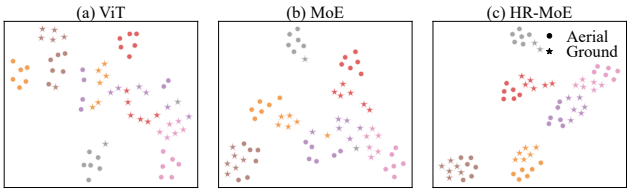

Figure 5: t-SNE visualization of extracted visual features. Each color represents a unique identity.

Effectiveness of view decoupling strategy. In the proposed view decoupling strategy, the view loss and orthogonal loss jointly facilitate viewpoint decoupling: the view loss explicitly guides the view token to capture viewpoint-specific features, while the orthogonal loss minimizes viewpoint information in the global feature. To assess their effectiveness, we conduct ablation studies by individually removing each loss, as shown in Table 4. It can be observed that (1) removing the orthogonal loss results in performance drop (No.4 *vs.* No.6), highlighting the negative effect of retaining viewpoint-specific information during image feature extraction and subsequent cross-modal alignment; (2) removing the view loss leads to a slight performance drop (No.5 *vs.* No.6), indicating that the model can still capture certain viewpoint-specific features even without explicit supervision. Additionally, Fig. 6 (left) shows that, the image-level router achieves high viewpoint classification accuracy, under the supervision of the view loss.

Ablation on viewpoints. To investigate the impact of diverse viewpoints, we evaluate TAG-CLIP on different viewpoint settings. Specifically, we select all $2,747$ identities with multi-view images from TAG-PEDES and randomly split them into a two groups (*i.e.*, $2,000$ IDs for training and 747 IDs for testing). Based on this, we construct three sub-datasets: one with only aerial-view images, one with only ground-view images, and one with a mix of aerial and ground views. All three datasets contain the same set of identities, differing only in the viewpoint of images. We then report the retrieval performance in Table 5. TAG-CLIP trained on multi-view data consistently outperforms its single-view counterparts across all evaluation settings (No.1 *vs.* No.2, No.3 *vs.* No.4, and No.5/No.6 *vs.* No.7), which stems from two factors: (1) the model's design effectively leverages multi-view information, and (2) multi-view training enhances feature robustness and generalization.

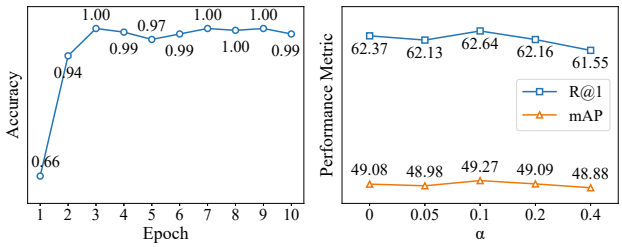

Figure 6: Left: Viewpoint classification accuracy of Image-level Router (Eq. 4). Right: Hyperparameter analysis of α.

Hyperparameter Analysis

The hyperparameter α in Eq. 10 controls the amount of viewpoint-specific information retained in the global visual feature. As shown in Fig. 6 (right), performance drops when α is too large (over-preserving view-specific cues) or too small (losing discriminative signals). At $\alpha = 0$, L_{ortho} becomes a standard orthogonal loss, which overly suppresses useful features. We set $\alpha = 0.1$ for trade-off. The analysis of λ_{id} and λ_{ortho} is provided in the Appendix.

Conclusion

In this paper, we introduce TAG-PR, which tackles person retrieval across a hybrid of aerial and ground image views, better reflecting real-world scenarios. To facilitate research in this task, we construct TAG-PEDES, which features diverse-view person images and high-quality synthetic textual descriptions. To address the challenges of heterogeneous image views, we propose TAG-CLIP, a method that incorporates the HR-MoE module and viewpoint decoupling strategy to enhance both image feature learning and cross-modal alignment. Extensive experiments demonstrate the effectiveness of TAG-CLIP. It is expected that our work can advance the field of person retrieval and drive its practical application in real-world scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 62476188 and 62176188, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China, Key Laboratory of New Generation Artificial Intelligence Technology & Its Interdisciplinary Applications (Southeast University), Ministry of Education, China.

References

Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023a. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv preprint arXiv:2305.13653*.

Bai, Y.; Ji, Y.; Cao, M.; Wang, J.; and Ye, M. 2025. Chat-based Person Retrieval via Dialogue-Refined Cross-Modal Alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3952–3962.

Bai, Y.; Wang, J.; Cao, M.; Chen, C.; Cao, Z.; Nie, L.; and Zhang, M. 2023b. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*, 757–767.

Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 465–473.

Cao, M.; Zhou, X.; Jiang, D.; Du, B.; Ye, M.; and Zhang, M. 2025. Multilingual Text-to-Image Person Retrieval via Bidirectional Relation Reasoning and Aligning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Y.; Huang, R.; Chang, H.; Tan, C.; Xue, T.; and Ma, B. 2021. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30: 4057–4069.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Fujii, T.; and Tarashima, S. 2023. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2786–2790.

Graves, A.; and Graves, A. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Jiang, J.; Ding, C.; Tan, W.; Wang, J.; Tao, J.; and Xu, X. 2025. Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9220–9230.

Jing, Y.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11189–11196.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, S.; He, C.; Xu, X.; Shen, F.; Yang, Y.; and Shen, H. T. 2024b. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3172–3180.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE international conference on computer vision*, 1890–1899.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017b. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; and Li, Z. 2021b. 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*, 16266–16275.

Liu, F.; Ye, M.; and Du, B. 2024. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2(1): 28.

Nguyen, H.; Nguyen, K.; Sridharan, S.; and Fookes, C. 2023. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2585–2590. IEEE.

Nguyen, H.; Nguyen, K.; Sridharan, S.; and Fookes, C. 2024. AG-ReID. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19: 2896–2908.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, 5566–5574.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, Z.; Hu, G.; and Zhao, C. 2024. Diverse person: Customize your own dataset for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4943–4951.

Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; and Tao, D. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17127–17137.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, 154–162.

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7297–7307.

Wang, S.; Wang, Y.; Wu, R.; Jiao, B.; Wang, W.; and Wang, P. 2025a. 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*, 22119–22128.

Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Wang, Y.; Yang, M.; Cao, R.; and Gao, G. 2025b. AEA-FIRM: Adaptive Elastic Alignment with Fine-Grained Representation Mining for Text-based Aerial Pedestrian Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XII 16*, 402–420. Springer.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32: 6032–6046.

Yan, S.; Liu, J.; Dong, N.; Zhang, L.; and Tang, J. 2024. Prototypical Prompting for Text-to-image Person Re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2331–2340.

Yang, S.; Zhou, Y.; Zheng, Z.; Wang, Y.; Zhu, L.; and Wu, Y. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4492–4501.

Zhang, Q.; Wang, L.; Patel, V. M.; Xie, X.; and Lai, J. 2024. 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*, 22000–22009.

Zhang, S.; Yang, Q.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; and Zhang, Y. 2023. Ground-to-aerial person search: Benchmark dataset and approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, 789–799.

Zhang, S.; Zhang, Q.; Yang, Y.; Wei, X.; Wang, P.; Jiao, B.; and Zhang, Y. 2020. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23: 281–291.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Bu, J.; and Tian, Q. 2015. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, 209–217.