Semantic-Driven Visual Progressive Refinement for Aerial-Ground Person ReID: A Challenging Large-Scale Benchmark

Aihua zheng[1,2,3], Hao Xie[1,3,4], Xixi Wan[2], Zi Wang[1,5], Shihao Li[2], Jin Tang[1,3,4], Bin Luo[1,3,4]*

[1]Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University
[2]School of Artificial Intelligence, Anhui University
[3]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation
[4]School of Computer Science and Technology, Anhui University
[5]School of Biomedical Engineering, Anhui Medical University
{ahzheng214, xhao2510,xixiwan11, ziwang1121, shli0603}@foxmail.com, {tangjin, luobin}@ahu.edu.cn

Abstract

Aerial-Ground Person Re-IDentification (AGPReID) aims to extract identity-discriminative representations from heterogeneous perspectives across different platforms in complex real-world environments. However, existing methods primarily focus on visual appearance modeling and make insufficient use of semantic attribute priors, which limits their ability to bridge the aerial-ground view gap. To address this limitation, we propose a Semantic-driven Visual Progressive Refinement framework for AGPReID (SVPR-ReID), which effectively leverages textual attribute priors to guide the extraction of fine-grained visual cues. Specifically, we design a View-Decoupled Feature Extractor that incorporates view-aware textual prompts to decouple view-invariant identity features. Then, to alleviate inter-class ambiguity, we propose an Attribute-Scattered Mixture-of-Experts module that integrates attribute semantics into the visual space, thereby improving discrimination among visually similar pedestrians. Finally, we design a Context-Vision Progressive Refinement module for progressive refinement of attribute and view-invariant features, obtaining robust cross-view identity representations. In particular, we contribute a comprehensive benchmark for AGPReID, named CP2108, which contains 142,817 images of 2,108 identities annotated with 22 attributes. Notably, it includes 191 identities captured across different times, enabling both short- and long-term ReID evaluation, addressing the limitation of existing datasets that focus only on short-term scenarios. Extensive experimental results validate the effectiveness of our SVPR-ReID on four AGPReID datasets.

Code — https://github.com/ahu-xhao/SVPR-ReID

Introduction

As a natural extension of Person Re-Identification (ReID), Aerial-Ground Person Re-IDentification (AGPReID) (Yan et al. 2021) focuses on matching identities across heterogeneous platforms, such as ground-level camera systems(Wang et al. 2022; Zheng et al. 2015; Wei et al. 2018; Zheng et al. 2021, 2023; Yan et al. 2023) and aerial-level camera systems (Wang et al. 2024; Chen, Ye, and Du 2022). However, the substantial domain gap between aerial and ground imagery results in severe viewpoint variations (Yan et al. 2021) and notable differences in resolution and imaging quality. These factors make cross-platform identity matching particularly challenging, as illustrated in Fig. 1. To address these challenges, existing AGPReID approaches can be broadly categorized into two paradigms. (1) Attribute-guided modeling: methods such as AGReID (Nguyen et al. 2023) and AGReID.v2 (Nguyen et al. 2024) leverage pedestrian attributes as semantic anchors for learning view-invariant local features, thereby alleviating cross-view matching difficulties. (2) Visual feature enhancement: VDT (Zhang et al. 2024) employs a hierarchical view-decoupling strategy within the ViT (Dosovitskiy et al. 2020) framework to extract view-invariant identity features. SeCap (Wang et al. 2025) enhances local vi-

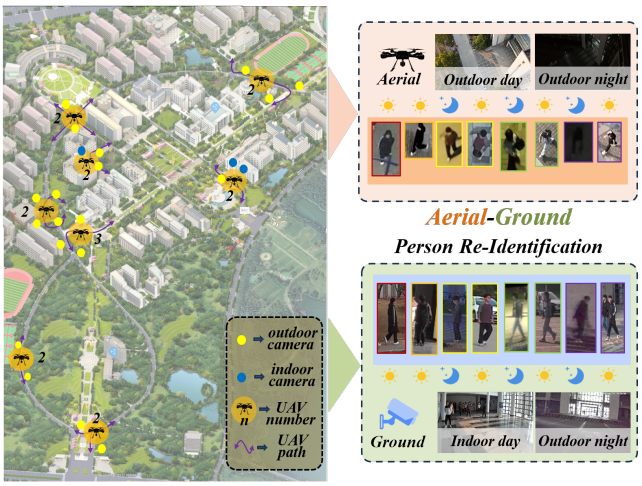


Figure 1: Aerial-Ground Person Re-IDentification poses significant challenges due to extreme variations in viewpoint (aerial vs. ground), scene context (indoor vs. outdoor), and time (day vs. night). Identities are annotated in the figure using bounding boxes of the same color.

*Corresponding author (luobin@ahu.edu.cn).

sual perception via adaptive visual prompts. Moreover, SD-ReID (Hu et al. 2025) leverages Stable Diffusion (Rombach et al. 2022) to generate auxiliary viewpoint representations. Although these methods are effective in mitigating cross-platform matching challenges, there are two key limitations: (1) insufficient exploitation of fine-grained semantic information contained in attribute cues (Zhang, Niu, and Zhang 2020), and (2) excessive reliance on local visual representations, leading to reduced robustness under adverse conditions such as low illumination (Zeng et al. 2020) or occlusion (He et al. 2021).

To address the limitations of existing approaches, we propose a Semantic-driven Visual Progressive Refinement framework for AGPReID (SVPR-ReID), which leverages textual attribute and viewpoint semantics to guide the learning of fine-grained, view-invariant identity representations. First, we introduce a View-Decoupled Feature Extractor (VDFE) to address CLIP's degradation under extreme viewpoint shifts (Nam et al. 2025). VDFE incorporates learnable view prompts to encode viewpoint semantics and decouple view-specific from view-invariant features, enhancing cross-view consistency while preserving identity cues. Secondly, to further tackle intra-class variance and inter-class ambiguity under occlusion and low-light conditions, we propose an Attribute-Scattered Mixture-of-Experts (ASMoE) module, which utilizes attribute priors and adaptive routing to dynamically integrate key attribute semantics. By complementing global features with fine-grained local cues, ASMoE effectively mitigates performance degradation due to viewpoint shifts and poor imaging conditions. Finally, we propose a Context-Vision Progressive Refinement (CVPR) module to ensure structural consistency based on image local patches. Unlike previous static local enhancement methods (He et al. 2021), CVPR adopts a multi-granularity attention mechanism to progressively refine the most discriminative patches, recovering subtle identity cues and reinforcing attribute modulation for robust cross-view representation.

Meanwhile, although several AGPReID datasets (Nguyen et al. 2023, 2024; Zhang et al. 2024; Wang et al. 2025) have been constructed to support research in this field, existing data resources remain inadequate. On one hand, some datasets (Zhang et al. 2024; Wang et al. 2025) lack essential semantic annotations such as attribute labels (Zhai et al. 2024), limiting the potential for attribute-guided (Kumar et al. 2021; Cho et al. 2022) semantic cues. On the other hand, despite simulating various illumination conditions, some datasets (Nguyen et al. 2023, 2024) fail to capture the same identities across different times of day (Li et al. 2024), limiting their applicability in long-term ReID scenarios (Nguyen et al. 2025).

To address the scarcity of AGPReID datasets, we contribute a large-scale real-world dataset CP2108 for comprehensive evaluation in the AGPReID task. The dataset contains 142,817 images of 2,108 pedestrian identities, captured by 22 ground cameras and 17 UAVs from two distinct models in eight diverse campus environments (such as library, crosswalks, indoor and outdoor teaching areas). CP2108 explicitly incorporates factors that significantly affect ReID performance, including seasonal variations, severe illumination changes, and occlusion. Notably, the dataset includes 191 identities enabling evaluation of long-term ReID capabilities, which is an aspect largely overlooked in prior datasets. Furthermore, each identity is annotated with 22 fine-grained pedestrian attributes (such as gender, hairstyle, clothing color, footwear type, carried objects, and body posture), thereby supporting research on semantic-guided alignment and attribute-based feature learning.

To summarize, this paper makes the following key contributions to the AGPReID task:

- We propose SVPR-ReID, a novel framework that integrates viewpoint and attribute semantics to achieve effective separation of viewpoint-invariant identity features and further enrich fine-grained feature representation.
- We design the CVPR module to exploit fine-grained visual cues and preserve structural consistency, thereby mitigating the degradation of local detail perception under low-quality conditions such as low illumination.
- We contribute a comprehensive AGPReID benchmark dataset named CP2108, which is annotated with extensive pedestrian attributes. It enables the exploration of diverse AGPReID scenarios, including long-term, multi-scene, and attribute-aware evaluation.
- Extensive experiments conducted on existing AGPReID benchmarks validate the effectiveness of our SVPR-ReID framework.

Related Work

Aerial-Ground Person Re-IDentification. The Aerial-Ground Person Re-IDentification (AGPReID) task presents significant challenges due to drastic appearance variations caused by viewpoint discrepancies from different platforms, which severely hinder reliable identity matching. To address these issues, representative AGPReID datasets have been constructed, including AGReID.v1 (Nguyen et al. 2023), AGReID.v2 (Nguyen et al. 2024), CARGO (Zhang et al. 2024) and LAGPeR (Wang et al. 2025). Based on these datasets, AGReID.v1 and AGReIDv2 employ pedestrian attributes as semantic anchors to facilitate visual alignment of local regions across aerial and ground views, enabling individuals with similar attributes to be more accurately associated, highlighting the critical role of local features in AGPReID. Another line of work focuses on learning robust view-invariant feature, as in VDT (Zhang et al. 2024), which designs a view-decoupled strategy based on the ViT (Dosovitskiy et al. 2020) architecture to hierarchically extract view-invariant identity representations.

Nowadays, prompt-based methods have shown promise in view-invariant modeling. SeCap (Wang et al. 2025), for instance, adaptively adjusts prompts (He et al. 2024) according to input images for different views, producing robust view-invariant representations. However, these methods are highly sensitive to image quality, and their stability degrades significantly under challenging conditions such as low illumination and motion blur. Unlike prior works, our SVPR-ReID fully exploits semantic information to guide

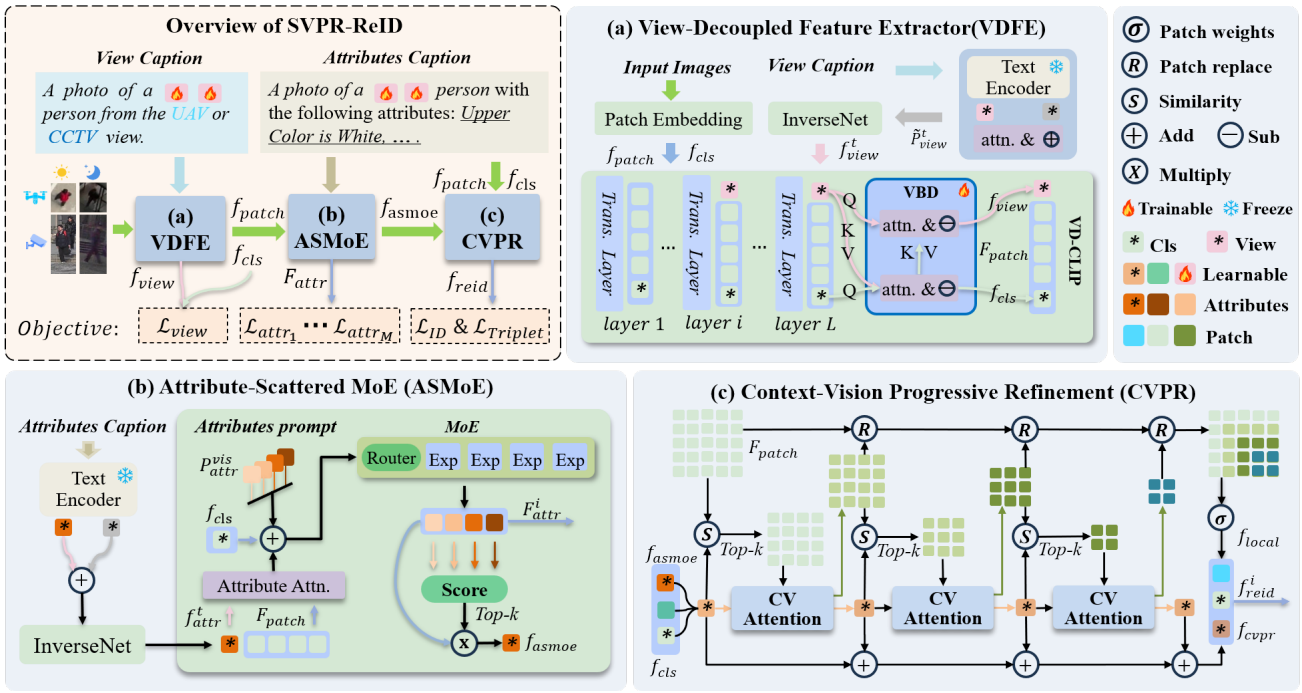


Figure 2: Overview of our SVPR-ReID framework. (a) View-Decoupled Feature Extractor (VDFE) encodes viewpoint semantics via learnable prompts to decouple view-specific and invariant features. (b) Attribute-Scattered Mixture-of-Experts (ASMoE) module adaptively integrates attribute semantics into identity features to enhance robustness under occlusion and low-light conditions. (c) Context-Vision Progressive Refinement (CVPR) module progressively refines discriminative patches based on attribute and identity cues through a multi-stage CVAttention mechanism for robust cross-view representation.

the model in effectively disentangling view-invariant features and mining fine-grained discriminative cues, thereby enhancing cross-view ReID performance.

Sematic-Driven Person Re-IDentification. Recently, CLIP's powerful image-text alignment capability (Radford et al. 2021) has inspired semantic (Jia, Chen, and Huang 2021) modeling in ReID tasks. Attributes (Gong, Liu, and Jain 2020), a critical semantic cue for distinguishing pedestrians (Zhang, Niu, and Zhang 2020), have been extensively utilized in two main ways. On the one hand, attribute label-based methods (Kumar et al. 2021; Huang et al. 2024; Wang et al. 2018b; Zhang et al. 2020) encode attribute annotations as supervision signals or auxiliary features to enhance visual representations. On the other hand, text-based approaches (Li, Sun, and Li 2023; Cui et al. 2024; Yu et al. 2024; Yang et al. 2024; Zhai et al. 2024) generate attribute descriptions for the samples and extract textual features for interaction with visual features. These studies demonstrate the potential of using textual priors (e.g., attributes) to model view-invariant features.

Methodology

As shown in Fig. 2, our SVPR-ReID framework consists of three key components: View-Decoupled Feature Extractor (VDFE), Attribute-Scattered Mixture-of-Experts (ASMoE) module, and Context-Vision Progressive Refinement (CVPR) module.

View-Decoupled Feature Extractor

We design VDFE based on CLIP (Radford et al. 2021) to adaptively decouple view-invariant identity features from viewpoint-related noise under the guidance of textual view prompts. Specifically, we design a viewpoint-sensitive template, *"A photo of a* $[x]_v$ $[x]_v$ $[x]_v$ $[x]_v$ *person from* ***aerial / ground*** *view."*, where the four consecutive $[x]_v$ are learnable view prompts P^t_{view}. The view descriptions are embedded into a global feature T_{view} via the frozen CLIP text encoder, and a residual cross-attention mechanism $\mathcal{CA}$ is then applied to obtain a high-level semantic view descriptor $\hat{P}^t_{view}$:

$$\hat{P}^t_{view} = P^t_{view} + \mathcal{CA}(P^t_{view}, T_{view}, T_{view}). \quad (1)$$

Next, we inverse $\hat{P}^t_{view}$ into the visual space to form a view-aware token $f^t_{view} = \varphi(\hat{P}^t_{view})$, where φ is a *MLP* layer. f^t_{view} is concatenated with the visual tokens f^i_{cls} and f^i_{patch} from the i-th CLIP layer and then fed into the remaining layers $\mathcal{V}_{L-i}$ to embed view semantics, forming:

$$\{f_{cls}, f_{patch}, f_{view}\} = \mathcal{V}_{L-i}(\{f^i_{cls}, f^i_{patch}, f^t_{view}\}), \quad (2)$$

where L denotes the number of CLIP layers.

Finally, we introduce a View-aware Bidirectional Decoupling (VBD) strategy to mitigate viewpoint interference, defined as follows:

$$f_{cls} = f_{cls} - \mathcal{CA}(f_{cls}, f_{view}, f_{view}), \quad (3)$$

$$f_{view} = f_{view} - \mathcal{CA}(f_{view}, f^i_{cls}, f^i_{cls}). \quad (4)$$

The resulting f_{cls} serves as the view-invariant identity feature, while f_{view} is discarded after training.

Attribute-Scattered Mixture-of-Experts

To reduce intra-class variance and inter-class ambiguity caused by occlusion or illumination changes, we design AS-MoE to dynamically inject attribute semantics into visual representations. Similar to the view-aware prompting strategy, we first define an attribute-aware textual template: *"A photo of a [x]$_a$ [x]$_a$ [x]$_a$ [x]$_a$ person with attributes: Pose is working, Upper Color is White, ..."*, where four *[x]$_a$* tokens form soft learnable attribute prompts P^t_{attr}. These descriptions are encoded as a global vector T_{attr} using the CLIP text encoder. We design a residual enhancement and a ***MLP*** inverseNet φ on P^t_{attr} and T_{attr} for a attribute token f_{attr}:

$$f^t_{attr} = \varphi(T_{attr} + \text{Mean}(P^t_{attr})). \tag{5}$$

Then, it is interacted with visual patches F_{patch} via cross-attention $\mathcal{CA}$ to produce a attribute context token $f_{attr} = \mathcal{CA}(f^t_{attr}, F_{patch}, F_{patch})$. Subsequently, a visual attribute prompts $\{P^{vis}_{attr,m}\}^M_{m=1}$ is processed by a ***MoE*** network conditioned on the f_{attr} and f_{cls}, formulated as:

$$F^i_{attr} = \{\textit{MoE}(f_{cls} + f_{attr} + P^{vis}_{attr,m})\}^M_{m=1}, \tag{6}$$

where M is the number of attributes and the ***MoE*** comprises four ***MLP*** experts and a linear gating network.

Considering some attributes may be unreliable under adverse conditions, we employ a top-k routing strategy $\mathcal{TK}$ to select the most discriminative attributes, which are aggregated with a learnable weight $\mathcal{W}$:

$$f_{asmoe} = \mathcal{W} \cdot \mathcal{TK}(\theta(F^i_{attr}), k), \tag{7}$$

where θ consist of a linear layer and a softmax operation.

Context-Vision Progressive Refinement

To further ensure structural consistency, we design the CVPR module as a fine-grained complement. It progressively enhances identity-relevant semantics by focusing on the most discriminative local patches, improving structural consistency across platforms. We sum the features f_{cls}, f_{asmoe}, and a learnable query to form the initial query f^1_{cvpr}, while the local patches F_{patch} are used as the key F^1_p. The CVPR applies a decoder CVAttention block, composed of self-attention, cross-attention, and feedforward layers. At refinement stage j of N, the process includes:

1) Top-k Patch Selection: Compute attention scores between f^j_{cvpr} and F^j_p to select top-k relevant patches $F^j_{p,k}$.

2) Feature Update: Use CVAttention on f^j_{cvpr} and $F^j_{p,k}$ to obtain refined query $f^j_{cvpr,r}$ and patches $F^j_{p,k,r}$, then update the global feature and selected patches with momentum γ:

$$\begin{aligned} f_{cvpr} &\leftarrow \gamma \cdot f^j_{cvpr} + (1-\gamma) \cdot f_{cls}, \\ F_{pp} &\leftarrow \gamma \cdot F^j_{pp} + (1-\gamma) \cdot F_{pp},\ \forall pp \in F^j_{p,k,r}. \end{aligned} \tag{8}$$

3) Iterative Refinement: The updated query and patches are propagated to the next stage for step 1) and 2) until $j = N$.

4) Final Aggregation: The refined token f_{cvpr} and the adaptively weighted local token $f_{local} = \bar{\mathcal{W}} \cdot F_{patch}$ are concatenated with the identity token f_{cls} to form the final representation f_{reid}, where $\bar{\mathcal{W}}$ denotes a learnable weight.

Optimization

As shown in Fig. 2, the training objective integrates three components: identity loss, view disentanglement loss, and attribute supervision. First, the Cross-Entropy loss and the Triplet loss are applied to the ReID feature f_{reid}:

$$\mathcal{L}_{reid} = \lambda_1 \mathcal{L}_{ID}(f_{reid}) + \lambda_2 \mathcal{L}_{Triplet}(f_{reid}), \tag{9}$$

where λ_1 and λ_2 are hyperparameters.

Second, the feature f_{view} and f_{cls} are regularized by a view classification and orthogonality constraint:

$$\mathcal{L}_{view} = \mathcal{L}_{ID}(f_{view}) + \sum_{j=1}^{|B|} |\langle f^j_{cls}, f^j_{view} \rangle|, \tag{10}$$

where $|\langle \cdot, \cdot \rangle|$ represents the absolute value after the dot product of two embeddings and $|B|$ denotes the batch size.

Finally, attributes are optimized via a Cross-Entropy loss:

$$\mathcal{L}_{attr} = \sum_{m=1}^{M} \Big(-\frac{1}{|B|} \sum_{i=1}^{|B|} c^m_i \log \hat{c}^m_i \Big), \tag{11}$$

where M is the number of attributes, while c_i represents the real label of the attribute and $\hat{c}_i$ denotes the prediction.

The overall objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{reid} + \alpha \mathcal{L}_{view} + \beta \mathcal{L}_{attr}, \tag{12}$$

where α and β are hyperparameters.

CP2108: A Comprehensive AGPReID Dataset

To further enrich the corpus of AGPReID task, we contribute a comprehensive AGPReID dataset, named CP2108.

Data Collection

The CP2108 dataset was collected over one calendar year on a university campus to comprehensively capture pedestrian appearances under real-world cross-platform conditions. It involves 2108 identities recorded by 17 UAVs and 22 static ground-based CCTVs. UAVs (DJI Mini4, Mavic 3T) followed ascending spiral flight paths across key campus regions (e.g., library, gates), capturing pedestrians from diverse altitudes (15–60m) and viewpoints (front, back, oblique). Ground cameras, installed at 2m height, provided complementary close-range views.

To enrich environmental diversity, the dataset was collected over a year, covering four seasons and multiple periods of the day. This design introduces illumination and scene variations such as day-light, backlighting, nighttime underexposure, and motion blur. Notably, 191 pedestrian identities were captured across day and night in different scenes, enabling robust evaluation of day-night matching for ReID.

Data Annotation

The annotation of the CP2108 dataset was conducted through a semi-automated workflow designed to balance efficiency with annotation accuracy. Initially, the YOLOv11 detector (Khanam and Hussain 2024) and the StrongSORT tracker (Du et al. 2023) were applied to all video streams

Property	AGReID	AGReID.v2	CARGO	G2APS-ReID	LAGPeR	CP2108 (Ours)
IDs	388	1615	**5000**	2788	4231	2108
Images	21983	100502	108563	**200864**	63842	142817
Source	real	real	Synthetic	real	real	**real**
Scenarios	1	1	1	1	7	**8**
Cameras	2 (1A+1G)	3 (1A+1G+1W)	13 (5A+8G)	2 (1A+1G)	21 (7A+14G)	**39 (17A+22G)**
Attributes	15	15	0	0	0	**22**
Captions	×	×	×	×	×	✓
Illumilation	Day	Day	Day & Night	Day or Night	Day or Night	**Day & Night**
Cross-Time	×	×	×	×	×	✓
Season	Spring, Summer	Spring, Summer	Virtual Summer	Summer	Summer	**All Year**
Altitude	15∼45m	1.5m, 3m, 15∼45m	5∼75m	20∼60m	20∼60m	**2m, 15∼60m**

Table 1: Statistics and key characteristics of Aerial-Ground Person ReID datasets, highlighting differences between CP2108 and prior benchmarks.

to automatically detect and track pedestrian targets. Following this initial automated stage, a team of 11 domain experts performed meticulous manual refinement of the annotations. This process involved correcting misaligned bounding boxes, supplementing missing detections, and removing redundant or low-quality images to ensure high consistency and reliability across diverse scenes and viewpoints.

To enhance semantic richness and support attribute-guided feature learning (Kumar et al. 2021), we annotate each identity with detailed attributes. Specifically, each pedestrian is labeled with 22 attributes grouped into physiological characteristics, appearance-related features, and environmental context.

Data Characteristic

As shown in Table 1 and Fig. 3, CP2108 extends existing AGPReID datasets by comprehensively covering five common challenges: viewpoint variation, elevative angles, partial occlusion, pose and motion blur, and heterogeneous illumination. Furthermore, it introduces three additional characteristics, namely multi-season variability, cross-time scenarios, and cross-scene scenarios.

1) Viewpoint and Elevation Diversity. CP2108 captures pedestrians from diverse perspectives, including front, back, side, top-down, and oblique views, supporting comprehensive modeling of cross-view variations. The UAVs adjust camera pitch angles from 30° to 85° during flight to produce multi-elevation data, and the image resolutions greatly range from 12×35 to 1366×1440 across platforms.

2) Occlusion and Motion Blur. The dataset features diverse occlusion scenarios, including umbrellas, trees, buildings, and dynamic crowds, causing partial visibility of key body parts. Motion blur arises from the object's movement and the UAV's rotation, further increasing visual complexity.

3) Illumination and Seasonal Variability. CP2108 captures pedestrian images at four times of day (morning, noon, evening, and night), covering diverse lighting conditions including overexposure and low illumination. It also spans four seasons, adding variations in clothing styles.

4) Cross-Time and Cross-Scene Diversity. A key feature of CP2108 is its support for long-term ReID and inter-scene

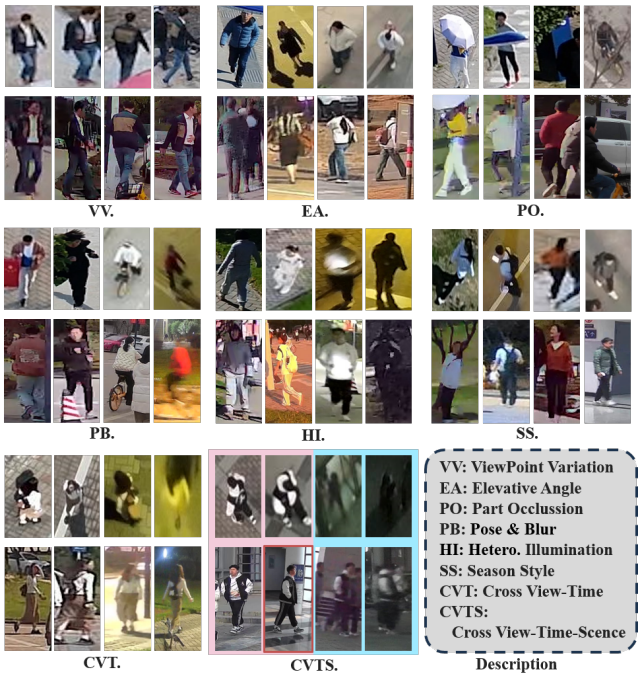


Figure 3: Visualization of key characteristics of our CP2108 dataset.

variation. We recruit 191 pedestrian identities for temporally controlled data collection and capture each identity under both daytime (Scene 1) and nighttime (Scene 2) conditions, enabling reliable day–night person retrieval. Additional indoor scenes further increase diversity and coverage.

Data Division

The CP2108 dataset contains 2108 identities and 142,817 images (approximately 67 images per identity, including 30 aerial and 37 ground views). We split the dataset into 1,442 IDs for training and 666 IDs for testing, including 191 cross-time identities, of which 71 are used for training and the rest for testing. Specifically, 436 IDs (109 cross-time) are

Method	AGReID.v1				CP2108 (Ours)					
	A→G		G→A		ALL		A→G		G→A	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
MGN (Wang et al. 2018a)	–	–	–	–	32.45	23.93	52.27	34.02	52.58	34.26
BoT (Luo et al. 2019)	70.01	55.47	71.20	58.83	36.27	28.35	57.76	38.60	55.64	37.72
SBS (He et al. 2023)	73.54	59.77	73.70	62.27	27.96	17.29	36.06	23.05	39.14	23.83
ViT (Dosovitskiy et al. 2020)	81.28	72.38	82.64	73.35	49.36	37.92	61.14	42.03	61.97	44.58
TransReID (He et al. 2021)	81.33	73.23	84.10	75.71	50.32	38.18	62.03	45.62	62.28	45.00
CLIP-ReID (Li, Sun, and Li 2023)	82.59	75.68	85.68	78.51	57.90	41.12	62.16	45.35	62.97	45.32
VDT (Zhang et al. 2024)	82.91	74.44	86.59	78.57	46.88	33.10	42.76	33.64	40.59	31.74
SeCap (Wang et al. 2025)	84.03	76.16	87.01	78.34	53.90	38.74	63.55	46.95	63.39	46.19
SVPR-ReID (Ours)	**85.34**	**77.85**	**87.32**	**80.55**	**59.23**	**42.40**	**64.25**	**47.72**	**63.76**	**47.04**

Table 2: Performance comparison on **AGReID.v1** and **CP2108 (Ours)** (in %) under multiple evaluation protocols. "A" and "G" denote aerial and ground views, respectively. "A→G" refers to aerial-to-ground matching and "G→A" refers to ground-to-aerial matching, while "ALL" indicates mixed retrieval. The best and second-best results are in bold and underlined, respectively.

randomly selected as queries, with all 666 IDs forming the gallery. Each query set includes one image per camera, while remaining images constitute the gallery. All splits are randomized for fairness and reproducibility.

Method	AGReID.v2		CARGO	
	A→C	W→A	ALL	A↔G
MGN (Wang et al. 2018a)	70.17	73.73	49.08	33.47
BoT (Luo et al. 2019)	77.03	75.90	46.49	32.56
ViT (Dosovitskiy et al. 2020)	77.03	76.59	53.54	40.11
TransReID (He et al. 2021)	78.93	72.96	53.17	38.06
CLIP-ReID (Li, Sun, and Li 2023)	83.27	80.56	62.08	52.54
VDT (Zhang et al. 2024)	79.13	78.52	55.20	42.76
SeCap (Wang et al. 2025)	80.84	80.15	60.19	58.94
SVPR-ReID (Ours)	**84.35**	**80.64**	**66.00**	**64.05**

Table 3: Comparison on AGReID.v2 and CARGO using mAP (%). "A", "G"/"C", and "W" denote aerial, ground, and wearable views, respectively.

Experiment

Experimental settings

Datasets. We conduct evaluations of SVPR-ReID on four AGPReID datasets: AGReID.v1 (Nguyen et al. 2023), AGReID.v2 (Nguyen et al. 2024), CARGO (Zhang et al. 2024), and our CP2108. ***AGReID.v1:*** This dataset consists of 21,893 images of 388 identities captured by ground and aerial cameras. It is divided into a training set with 199 IDs and a test set with 189 IDs. Each identity is annotated with 15 attributes. We follow the original protocol and report results under two protocols: A→G and G→A. ***AGReID.v2:*** An extended version of AGReID.v1, containing 1,615 identities captured by UAV, ground, and wearable cameras. It is split into 807 identities for training and 798 for testing. Evaluation follows two protocols: A→C and W→A. ***CARGO:*** A synthetic dataset with 5,000 identities captured by 8 ground and 5 aerial cameras. It is split into 2,500 identities for training and 2,500 for testing, using two protocols: ALL and A↔G. ***CP2108:*** The training set contains 1,442 identities, and the remaining 666 identities form the test set. Each identity is annotated with 22 attributes. To comprehensively evaluate cross-view performance, we adopt three protocols: ALL, A→G, and G→A, following common practice.

Evaluation Metrics. We evaluate retrieval performance using Rank-1 accuracy and mean Average Precision (mAP). These metrics jointly assess retrieval accuracy and ranking quality, providing an overall measure of model performance under challenging cross-view conditions.

Implementation Details. The experiments are conducted using PyTorch on an NVIDIA GTX 4090 GPU. We adopt CLIP, specifically the CLIP-Base-16 (Li, Sun, and Li 2023), as the backbone network. During both training and inference, all input images are resized to 256×128 for all AGPReID datasets. The loss weights λ_1, λ_2, α, and β are set to 0.25, 1.0, 1.0, and 0.1, respectively. The momentum γ is set to 0.1, and the optimizer is Adam with a learning rate of 5×10^{-6}. The batchsize is 128 sampled from 32 identities.

Comparison with State-of-the-Art Methods

Comparison on CP2108. As shown in Table 2, our SVPR-ReID significantly outperforms single-view and cross-view baselines under **ALL** evaluation protocols. Specially, SVPR-ReID surpasses the strong cross-view method SeCap (Wang et al. 2025) by a substantial margin of **5.33%** in Rank-1 and **3.66%** in mAP. These gains validate the effectiveness of our dual-semantic guidance strategy, which combines viewpoint-aware and attribute-aware prompts with progressive refinement to generate robust and discriminative AGPReID representations.

Comparison on AGReID.v1. Similarly, we evaluate our method on the AGReID.v1 dataset to further verify its generalization capability. As shown in Table 2, compared with the strong single-view method CLIP-ReID (Li, Sun, and Li 2023), our SVPR-ReID yields significant improvements of **2.75%** in Rank-1 and **2.17%** in mAP on the A→G protocol. These results highlight the superior ability of our SVPR-ReID to handle severe viewpoint changes and complex real-world conditions.

Comparison on AgReID.v2 and CARGO. As shown in

Table 3, our SVPR-ReID demonstrates strong performance on both AGReID.v2 and CARGO. Specifically, on CARGO, despite the absence of attribute labels, it outperforms SeCap by 5.11% in mAP on the $A \leftrightarrow G$ protocol. These results confirm the robust generalization of our SVPR-ReID.

Comparison on CP2108 for Cross-Time scenarios. To further assess robustness under illuminations, we conduct experiments on a cross-time setting of CP2108, as shown in Table 4. Our SVPR-ReID improves Rank-1 by 3.34% and mAP by 3.51% over the baseline under the **ALL** protocol, demonstrating the role of attribute-aware guidance in challenging long-term scenarios. Here, we use CLIP-ReID without the text encoder as the baseline. However, the performance remains slightly below CLIP-ReID, likely due to large appearance changes from clothing variations across days (Pathak and Rawat 2025), which reduce the reliability of attribute cues. This highlights a key limitation and calls for more adaptive attribute modeling strategies to handle clothing variations and illumination shifts in future work.

Method	ALL		A ↔ G	
	Rank-1	mAP	Rank-1	mAP
TransReID (He et al. 2021)	40.37	28.77	34.77	27.47
CLIP-ReID (Li, Sun, and Li 2023)	43.41	33.15	**36.91**	**33.78**
VDT (Zhang et al. 2024)	40.61	29.78	35.36	28.35
SeCap (Wang et al. 2025)	44.01	33.96	35.62	30.83
Baseline	41.07	30.98	35.12	29.68
SVPR-ReID (Ours)	**44.41**	**34.49**	35.94	31.13

Table 4: Performance comparison with state-of-the-art methods under two protocols on the CP2108 dataset for **cross-time** scenario.

Ablation Study

Effects of Key Components. As shown in Table 5, under the ALL protocol, the VDFE yields the largest single-strategy improvement boosting Rank-1 by 1.9% by mitigating view-specific noise.. Adding the CVPR brings an additional gain of 1.7% in Rank-1, higher than that of the ASMoE, indicating that preserving structural consistency is more critical for AGPReID. When all strategies are combined, our method achieves the best overall performance with a Rank-1 of 59.23% and mAP of 42.40%, confirming the complementary benefits of textual semantic embedding and local details for robust cross-view representations.

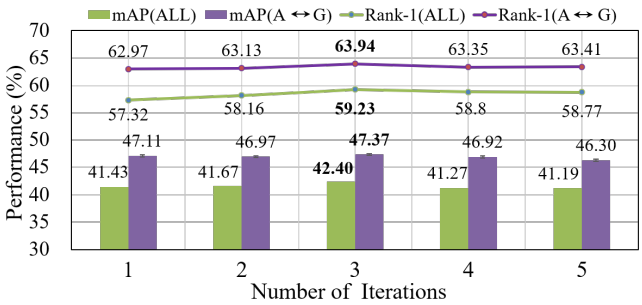
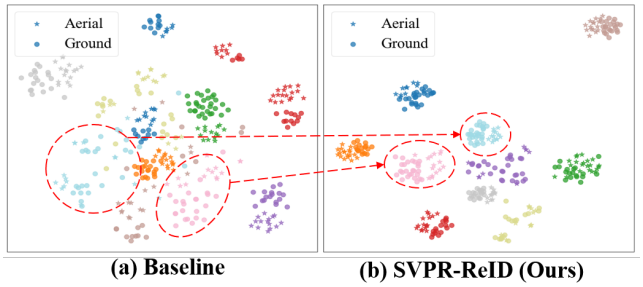


Figure 4: Ablation of the number of iterations in CVPR.

Discussion on CVPR. We evaluate the impact of different iteration numbers on performance. As shown in Fig. 4, three iterations achieve the best results with 42.40% mAP and 59.23% Rank-1 under the ALL protocol. Increasing the iterations beyond three yields only marginal gains or slight drops while increasing additional computational cost.

Component	ALL		A↔G	
	Rank-1	mAP	Rank-1	mAP
Baseline	56.15	39.97	61.42	45.21
+VDFE	57.10	40.85	62.38	46.20
+VDFE+ASMoE	58.05	41.50	63.00	46.75
+VDFE+CVPR	58.80	42.10	63.43	47.00
+VDFE+ASMoE+CVPR	**59.23**	**42.40**	**63.94**	**47.37**

Table 5: Ablation study of key components in SVPR-ReID on the CP2108 dataset.



(a) Baseline **(b) SVPR-ReID (Ours)**

Figure 5: Visualization of feature distributions with t-SNE. Different colors refer to different IDs, while different shapes refer to different views.

Visualization

As shown in Fig. 5, the t-SNE visualization further illustrates that our method produces more compact intra-class clusters and clearer inter-class separation. It highlights the importance of incorporating additional semantic guidance for viewpoint disentanglement and progressive feature refinement in addressing the challenges of AGPReID.

Conclusion

This paper explores the challenging task of AGPReID characterized by viewpoint changes and illumination conditions. We propose **SVPR-ReID**, a novel framework that leverages viewpoint and attribute priors for robust cross-view representation. Specifically, VDFE disentangles view-specific and view-invariant features, ASMoE dynamically injects attribute semantics to mitigate ambiguity, and CVPR progressively refines local patches to recover fine-grained identity cues under occlusion or illumination degradation. To further advance AGPReID research, we contribute **CP2108**, a comprehensive real-world dataset with 142K images of 2,108 identities, including 191 identities spanning day-night scenarios. Each identity is annotated with 22 fine-grained attributes. However, our method still faces limitations in long-term ReID scenarios involving clothing changes, which will be an important direction for future exploration.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62372003 and 62576006, the Natural Science Foundation of Anhui Province under Grants 2308085Y40, and the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University (Grant No. 2024A004).

References

Chen, S.; Ye, M.; and Du, B. 2022. Rotation invariant transformer for recognizing object in uavs. In *Proceedings of the ACM International Conference on Multimedia*, 2565–2574.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022. Part-based Pseudo Label Refinement for Unsupervised Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7318.

Cui, C.; Huang, S.; Song, W.; Ding, P.; Zhang, M.; and Wang, D. 2024. Profd: Prompt-guided feature disentangling for occluded person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 1583–1592.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; and Meng, H. 2023. Strongsort: Make deepsort great again. *In IEEE Transactions on Multimedia*, 25: 8725–8737.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *Proceedings of the European conference on computer vision*, 330–347.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2023. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 9664–9667.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

He, W.; Deng, Y.; Tang, S.; Chen, Q.; Xie, Q.; Wang, Y.; Bai, L.; Zhu, F.; Zhao, R.; Ouyang, W.; et al. 2024. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17521–17531.

Hu, X.; Zhang, P.; Wang, Y.; Yan, B.; and Lu, H. 2025. SD-ReID: View-aware Stable Diffusion for Aerial-Ground Person Re-Identification. *arXiv preprint arXiv:2504.09549*.

Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; and Wang, L. 2024. Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17689–17699.

Jia, J.; Chen, X.; and Huang, K. 2021. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 962–971.

Khanam, R.; and Hussain, M. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Kumar, S. V. A.; Yaghoubi, E.; Das, A.; Harish, B. S.; and Proença, H. 2021. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices. *IEEE Transactions on Information Forensics and Security*, 16: 1696–1708.

Li, H.; Chen, J.; Zheng, A.; Wu, Y.; and Luo, Y. 2024. Day-night cross-domain vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12626–12635.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1405–1413.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Nam, J.; Im, J.; Kim, W.; and Kil, T. 2025. Extract free dense misalignment from CLIP. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6173–6181.

Nguyen, H.; Nguyen, K.; Pemasiri, A.; Liu, F.; Sridharan, S.; and Fookes, C. 2025. AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1241–1251.

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

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.

Pathak, P.; and Rawat, Y. S. 2025. Colors see colors ignore: Clothes changing reid with color disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16797–16807.

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 *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018a. Learning discriminative features with multiple granularities

for person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, 274–282.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018b. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2275–2284.

Wang, L.; Zhang, Q.; Qiu, J.; and Lai, J. 2024. Rotation exploration transformer for aerial person re-identification. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 1–6.

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

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, 2633–2641.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 79–88.

Yan, C.; Pang, G.; Wang, L.; Jiao, J.; Feng, X.; Shen, C.; and Li, J. 2021. BV-person: A large-scale dataset for bird-view person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10943–10952.

Yan, P.; Liu, X.; Zhang, P.; and Lu, H. 2023. Learning convolutional multi-level transformers for image-based person re-identification. *Visual Intelligence*, 1: 24.

Yang, Z.; Wu, D.; Wu, C.; Lin, Z.; Gu, J.; and Wang, W. 2024. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17343–17353.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6764–6772.

Zeng, Z.; Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2020. Illumination-adaptive person re-identification. *IEEE Transactions on Multimedia*, 22: 3064–3074.

Zhai, Y.; Zeng, Y.; Huang, Z.; Qin, Z.; Jin, X.; and Cao, D. 2024. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6979–6987.

Zhang, J.; Niu, L.; and Zhang, L. 2020. Person re-identification with reinforced attribute attention selection. *IEEE Transactions on Image Processing*, 30: 603–616.

Zhang, Q.; Wang, L.; Patel, V. M.; Xie, X.; and Lai, J.-H. 2024. View-decoupled Transformer for Person Re-identification under Aerial-ground Camera Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22000–22009.

Zhang, S.; Chen, C.; Song, W.; and Gan, Z. 2020. Deep feature learning with attributes for cross-modality person re-identification. *Journal of Electronic Imaging*, 29(3): 033017–033017.

Zheng, A.; Liu, J.; Wang, Z.; Huang, L.; Li, C.; and Yin, B. 2023. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1(1): 29.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust Multi-Modality Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4, 3529–3537.

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*.