AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification

Supplementary Material

8. Dataset

8.1. Soft-biometric Attributes

Following [31, 32], a comprehensive framework of 15 soft biometric attributes (Fig. 5) is employed to facilitate cross-view person identification. These attributes are categorized as physical traits (gender, age, height, weight, ethnicity, hairstyle, beard, and mustache) and appearance traits (glasses, head accessories, upper/lower body clothing, footwear, and accessories). This selection builds upon prior surveillance research [17, 22], prioritizing characteristics that are both discriminative and consistent across aerial and ground viewpoints for practical real-world challenges. The distribution of these soft-biometric attributes is illustrated in Fig. 6.

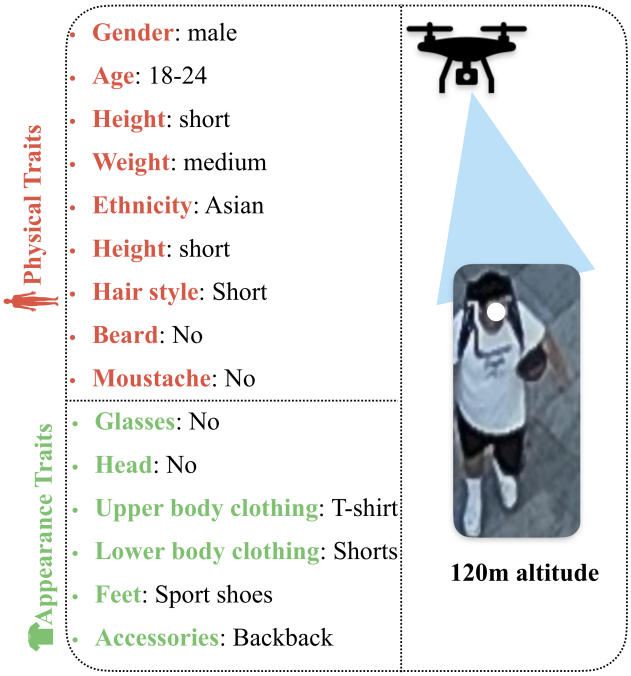

Figure 5. Soft-biometric attributes in our AG-VPReID dataset, showing Physical (top) and Appearance (bottom) Traits of a person from a top view. The attributes are categorized into physical characteristics (such as gender, age, height) and appearance details (such as clothing and accessories).

8.2. Long-term Re-identification

To capture realistic long-term appearance variations, we recruited 14 volunteers (consisting of nine males and five females) to participate in our data collection over a period of eight weeks with non-consecutive recording sessions. Each participant attended multiple recording sessions, with an average of four sessions per person, deliberately changing their attire between sessions. The clothing changes included variations in style (e.g., formal wear, casual wear, athletic wear), color schemes, and outer layers (e.g., jackets, coats). We instructed participants to wear clothing from their personal wardrobes to ensure naturalistic appearance variations. The recording sessions were scheduled at different times of day and under varying weather conditions, adding environmental diversity to our dataset. Each participant's sessions were separated by a minimum interval of 14 days to maximize clothing variation and capture realistic long-term appearance changes.

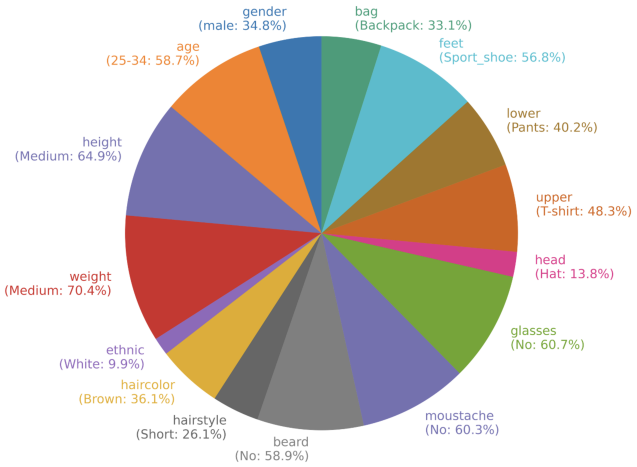

Figure 6. Most common soft-biometric attributes in our dataset.

8.3. Calibration

The intrinsic camera parameters, including focal length (f_x, f_y), principal point (c_x, c_y), distortion coefficients (k_1, k_2, p_1, p_2), and GPS coordinates, are presented in Tab. 8. Relative camera locations and viewing angles are visualized in Fig. 1 of the main paper. The camera models, resolutions, lenses, and frame rates are listed in Tab. 2 of the main paper.

9. Approach

9.1. AG-VPReID-Net Framework Overview

The AG-VPReID-Net framework, as detailed in Table 9, provides a comprehensive approach to aerial-ground video

Camera	f_x	f_y	c_x	c_y	k_1	k_2	p_1	p_2	Location
Bosch Outdoor	16123.85	16123.85	298.39	425.13	-0.89	1.06	0.73	1.79	(27°28'36"S, 153°01'45"E)
Bosch Indoor	17129.38	17129.38	306.20	424.81	-1.27	1.27	0.71	2.06	(27°28'40"S, 153°01'44"E)
GoPro10	17141.64	17141.64	303.62	401.73	-1.66	2.15	0.48	1.61	(27°28'35"S, 153°01'44"E)
GoPro10	16843.23	16843.23	264.75	428.89	-1.40	2.03	0.14	1.26	(27°28'35"S, 153°01'44"E)
DJI Inspire2	16467.97	16467.97	302.48	372.46	-0.93	1.14	0.41	1.68	(27°28'37"S, 153°01'46"E)
DJI M300RTK	16837.10	16837.10	291.64	418.17	-0.94	1.20	0.38	1.96	(27°28'40"S, 153°01'46"E)

Table 8. Intrinsic camera parameters and GPS coordinates

Stream	**Focus**	**Challenges Addressed**
1. Adapted Temporal-Spatial Stream	• Temporal shape modeling • Identity-specific memory • Pre-trained large vision models	• Motion pattern inconsistencies • Temporal discontinuity • Sequential feature learning
2. Normalized Appearance Stream	• UV map-based appearance normalization • Physics-informed techniques • 3D appearance representation	• Drastic changes in resolution • Appearance variations between aerial and ground views • Pose variations and partial occlusions
3. Multi-Scale Attention Stream	• Multi-scale feature extraction • Transformer decoder • Local temporal module	• Scale variations due to varying drone altitudes • Integration of spatial and temporal information • Fine-grained detail capture
Overall Framework	• Combination of all three streams • Cross-platform visual-semantic alignment	• Robust person representation across different views • Addresses viewpoint, resolution, scale, and occlusion

Table 9. Overview of AG-VPReID-Net streams and their contributions

person re-identification. This framework comprises three specialized streams, each designed to address specific challenges. The Adapted Temporal-Spatial Stream focuses on motion patterns and sequential feature learning, while the Normalized Appearance Stream employs UV map-based normalization and 3D appearance representation. The Multi-Scale Attention Stream utilizes multi-scale feature extraction and a transformer decoder. These streams combine to form a robust solution. This solution handles the complex challenges of matching individuals across aerial and ground-based video footage, including viewpoint variations, occlusions, and scale differences.

9.2. Optimization

Our AG-VPReID-Net achieves optimal performance by integrating three streams. Initially, each stream produces independent feature representations. These are then combined using an adaptive weighted fusion mechanism:

$$F_{combined} = \alpha F_{temporal} + \beta F_{appearance} + \gamma F_{multiscale}, \tag{8}$$

where α, β, and γ are learnable parameters that adapt to the input characteristics. This allows the model to dynamically emphasize different streams depending on the specific aerial-ground matching scenario.

To further improve ranking performance, we employ Reciprocal Rank Fusion (RRF) as a post-processing step. RRF combines the individual rankings from each stream to produce a final, more robust ranking:

$$RRF(d) = \sum_{i=1}^{3} \frac{1}{k + r_i(d)}, \tag{9}$$

where d is a candidate match, k is a constant set to 60 in

our experiments, and $r_i(d)$ is the rank of d in the i-th stream. This fusion technique gives higher weight to candidates that rank highly across multiple streams, leading to improved mean Average Precision (mAP) and Rank-k accuracy in our experiments.

RRF is chosen for its robustness, handling cases where correct matches may have inconsistent rankings across streams. This is particularly important in aerial-ground ReID, where different streams may excel under different conditions (e.g., varying altitudes or occlusions). By combining rankings rather than raw scores, RRF provides a more robust final ranking that is less sensitive to individual stream failures.

10. Implementation Details

UV Map Acquisition and Processing. Our implementation pipeline uses UV maps from Texformer [41], which utilizes 3D human meshes from RSC-Net [42]. To preserve temporal relationships within individual video sequences, we integrate PhysPT [49] for more precise pose estimations. These refined poses are fed into Texformer, yielding higher-fidelity UV maps. We enhance inter-frame consistency through normalization, histogram matching, and gamma correction. The final UV map is constructed via weighted blending, combining processed UV maps with a visibility mask M = max(dot(N, V), 0), where N is the surface normal and V is the view vector. Blending weights are determined using a softmax function over mask values, ensuring smooth transitions between UV map regions.

Configuration for Stream 1: Adapted Temporal-Spatial Stream. The Adapted Temporal-Spatial Stream utilizes a pre-trained CLIP ViT-B/16 model as the visual encoder (frozen during training), complemented by a Temporal Shape Modeling (TSM) branch with 2 GRU layers (1024 neurons each) and an identity-aware 3D regressor. An Attention-based Shape Aggregation (ASA) module, consisting of 2 GRU layers and a self-attention mechanism, processes shape information. Temporal features are captured using a Temporal Memory Diffusion (TMD) module with multi-head self-attention. The stream is trained on 8-frame clips (256×128 resolution) with a batch size of 16, using Adam optimizer, an initial learning rate of 5×10^{-3} with warm-up and decay, and a weight decay of 0.01. Data augmentation includes random horizontal flipping and random erasing.

Configuration for Stream 2: Normalized Appearance Stream. The Normalized Appearance Stream processes 3D coordinates and normalized UV texture through four Omni-scale Modules. Each module incorporates a UV-space adapted Dynamic Graph Convolution (DGC) layer and three parallel branches with varying grouping rates. The network transforms the initial $m \times 6$ input (UV coordinates + RGB) into a 96×512 feature map, which is then reduced to a 512-dimensional feature vector via global pooling and a fully connected layer. The training utilizes cross-entropy loss for identity classification, with Adam optimization and cosine learning rate scheduling over 1000 epochs. During inference, the final 512-dimensional feature vector serves as a robust representation of normalized appearance across multiple frames.

Configuration for Stream 3: Multi-Scale Attention Stream. The CLIP ViT-L/14 vision encoder is employed to extract multi-scale features, utilizing a Pad-and-Resize technique to uniformly adjust each frame to a resolution of 224×224, thereby preserving the original body proportions. The resulting feature volume, $\mathbf{G} \in \mathbb{R}^{T \times 257 \times d}$, encapsulates the embedding size of the token, d. Following the methods [23, 44], the feature maps from the last four layers of the image encoder (i.e., $N = 4$) are utilized. Additionally, four Transformer decoder blocks ($M = 4$) are applied to further process these features.

Figure 7. A2G: Comparing baseline vs our method on AG-VPReID. Green/red: correct/incorrect labels. Altitudes: 15m (orange), 120m (blue). First tracklet image is shown. Ranks show improvements in **bold**. Best in color.

11. Visualization

To provide a comprehensive analysis of our method's effectiveness compared to the baseline CLIP-based approach [21], we present qualitative results through visual-

ization examples in Fig. 7 and Fig. 8. For the aerial-to-ground (A2G) matching scenario shown in Fig. 7, we compare the retrieval results between our approach and the baseline method. Each query image is captured from aerial views at different altitudes (15m and 120m), with corresponding ground-truth matches from ground-level perspectives. The green and red boxes indicate correct and incorrect matches respectively, while improved rank positions are highlighted in bold. These visualizations demonstrate our method's superior ability to handle severe viewpoint variations and maintain reliable person matching across aerial and ground views. Additional examples for ground-to-aerial (G2A) matching presented in Fig. 8 further validate the robustness of our approach.

Figure 8. G2A: An extended case of Fig. 7