KFUPM
Journal of Undergraduate Research International

Diamond Open Access 48

Article Computer Science

Zero-Shot UAV Person Re-ID via CLIP Embeddings and Soft Biometrics

Lama Al-Thunayyan and Hussein Samma*

Cite https://doi.org/10.64589/juri/231614

Submitted: July 31, 2025 **Revised:** April 9, 2026 **Accepted:** April 21, 2026

ABSTRACT

Person re-identification (Re-ID) in unmanned aerial vehicle (UAV)-captured images presents unique challenges due to extreme variations in viewpoint, low image resolution, and frequent occlusions. This study develops and evaluates a zero-shot Re-ID framework that leverages contrastive language-image pre-training (CLIP)-based visual embeddings fused with soft biometric attributes, enabling person retrieval without requiring training data. This approach extracts interpretable semantic features such as clothing color, gender, and body proportions, and combines them with high-level vision-language representations generated by CLIP to enable robust identity encoding. By integrating visual and textual modalities, the framework supports flexible query-based person searches for identities not seen during training. Experimental evaluation and comparative analysis demonstrate the effectiveness of the method, achieving a Rank-1 accuracy of 75.87% on the Aerial-Ground Re-ID (AG-ReID.v2) dataset, which comprises person Re-ID across mixed aerial and ground scenarios, surpassing other deep-learning approaches under zero-shot conditions. These results confirm that combining soft biometrics with vision-language models enables scalable, training-free person re-identification in real-world aerial surveillance applications.

Keywords: person re-identification (Re-ID), CLIP, soft biometrics, zero-shot, AG-ReID.v2 dataset

1. INTRODUCTION

Person re-identification (Re-ID) in unmanned aerial vehicle (UAV)-captured images used for aerial surveillance, search-and-rescue operations, and intelligent monitoring systems is a critical challenge in computer vision. Compared to conventional ground-based Re-ID, UAV scenarios involve extreme variations in viewpoint, low image resolution, frequent occlusions, and dynamic backgrounds, which severely hinder model robustness and accuracy. These challenges are illustrated in Figure 1, where the same individual is captured from two UAV viewpoints with noticeable differences in viewing angle, walking posture, body orientation, and lighting conditions. Such changes, even for the same person, can significantly affect the appearance matching and complicate the re-identification process.

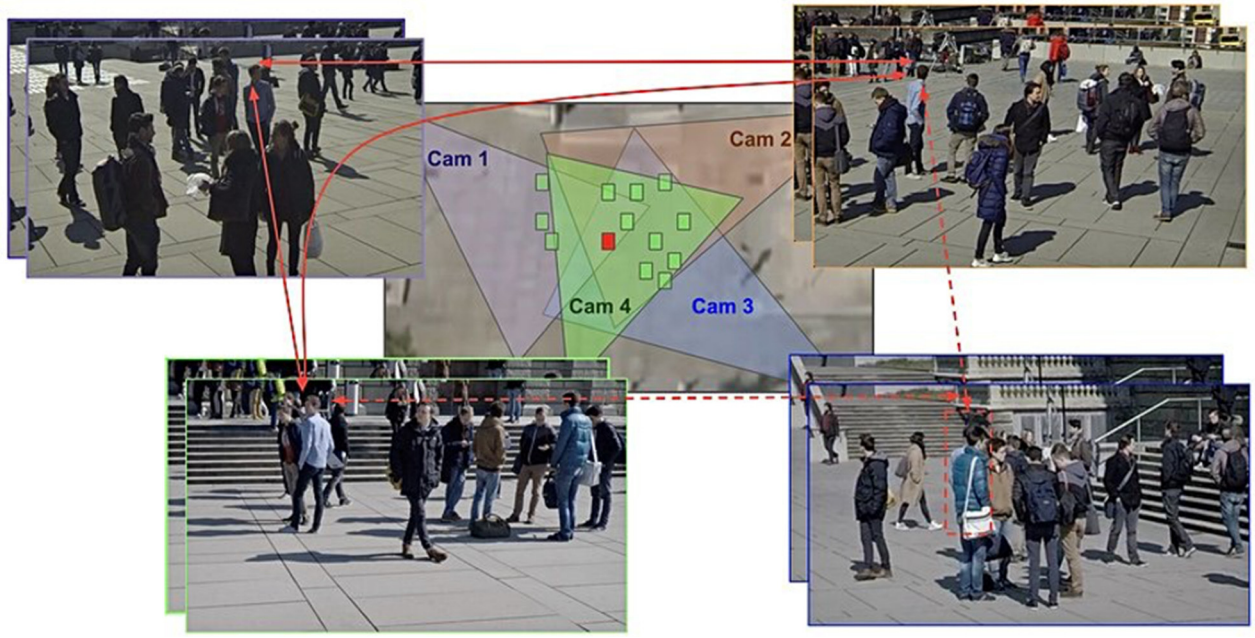
Traditional deep-learning approaches typically employ convolutional neural networks (CNNs) or transformer-based architectures trained on annotated person datasets. Although effective under controlled settings, these models suffer from limited generalization in zero-shot environments where unseen identities are encountered. In addition, domain-adaptation methods require additional training and labeled target-domain data, making them unsuitable for real-time UAV applications. Previous studies on person Re-ID have explored various model architectures, focusing on aerial or cross-modality settings. Transformer-based models, such as vision transformers (ViTs)[1], have shown strong performance under wide-view variations, but require large-scale labeled datasets for effective generalization. The convolutional part refinement (CPR) model[2] achieves a high Rank-1 accuracy on benchmark datasets by employing multichannel feature refinement and label smoothing techniques. However, the use of dual backbones introduces high computational overhead, which limits real-time deployment. Similarly, the contrastive pedestrian attentive and correlation learning (CpaCol) model[3] addresses occlusions using attention and correlation modules; however, it depends heavily on large and diverse datasets. In the cross-modality domain, the CLIP-driven Semantic Discovery Network (CSDN)[4] demonstrates a strong alignment between visual and infrared features using dual text prompts and fusion mechanisms, but at the cost of increased system complexity.

The dual-attention biometric-clothes transfusion network (DualBCT-Net)[5] presents a combined strategy that leverages both biometric information and a clothing template, which is an additional image representing the outfit the person is expected to wear. These features are extracted through separate branches and fused using a dual attention mechanism. Although effective, this method assumes access to a reliable clothing template, which may not be feasible in real-world scenarios. The multimodal-assisted feature learning framework (MAL-F)[6] integrates RGB, grayscale, and contour images to learn representations that are robust to clothing changes, enabling re-identification across different outfits. Combined with the Residual-Transformer Network (ResT-Net) and the Non-Clothing Salient Region (NC-SR) module, the method suppresses clothing-related cues and emphasizes shape and structure. Although it achieves strong performance,

Published under an exclusive license by KFUPM



(a) Same individual captured from two different viewpoints.



(b) Target captured by different cameras with differences in lighting, and one image is occluded.

Figure 1. Challenges in UAV-based person re-identification

the training process remains relatively demanding compared with more lightweight alternatives, such as our zero-shot CLIP-based approach. The transformer-based object Re-ID framework (TransRL-ReID)[7] employs a vision transformer with two core components: pixel swapping to suppress clothing features and a DropAll module that regularizes attention weights. This enables the model to focus on identity cues, such as body shape and silhouette. However, this method depends on human parsing to segment clothing regions before training, which may introduce noise and affect the feature quality if the parsing is inaccurate.

Most recently, several studies on UAV-based person Re-ID have focused on improving cross-view robustness and real-world applicability. For instance, Nguyen and Mai[8] proposed a CLIP-based aerial–ground Re-ID method with hard sample mining, achieving improved Rank-1 accuracy and robustness on video datasets. The DetReIDX benchmark introduced a large-scale UAV dataset (13M+ boxes and 509 IDs), showing that existing methods suffer significant performance drops and highlighting real-world challenges. Additionally, Hambarde et al.[9] explored long-range UAV Re-ID using video modeling and multi-view

fusion, improving Re-ID performance under extreme conditions, although generalizability remained limited.

Despite these advances, most existing Re-ID methods depend on supervised training and domain-specific optimization, which limit their adaptability to dynamic UAV environments. To address these limitations, we propose a zero-shot UAV person Re-ID framework that leverages CLIP embeddings and soft biometrics to enable training-free retrieval. Our approach extracts semantic attributes, such as clothing color, gender, and body proportions, and fuses them with CLIP-generated visual embeddings to construct a robust identity representation. This fusion enables the system to perform an accurate person search even when individuals have never been seen before. The main contributions of this study are as follows:

1. Zero-Shot Re-ID with CLIP: This approach utilizes CLIP's image-text alignment capabilities to retrieve individuals based on textual descriptions, eliminating the need for supervised fine-tuning.
2. Soft Biometrics Integration: Enhances retrieval accuracy by incorporating human-interpretable attributes, which are particularly valuable in low-resolution occluded aerial footage.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology, Section 3 discusses the experimental analysis, and Section 4 concludes the paper with a summary and directions for future work.

2. METHODOLOGY

Figure 2 illustrates the main framework of the proposed Zero-Shot UAV Person Re-ID approach. The pipeline comprises several stages: UAV image input, person-region extraction, feature extraction, similarity fusion, and final matching. Each of these stages is described below.

Table 1. *AG-ReID.v2 Dataset properties*

Property	Value
Total Identities	1,615
Query Images	4,348
Gallery Images	21,214
Image Resolution	Varies
Viewpoint Range	15–45 m

2.1. UAV Input Image. In this study, we used UAV images from the AG-ReID.v2 dataset. The AG-ReID.v2 dataset comprises 4,348 query images and 21,214 gallery images of 1,615 unique individuals (Table 1). The varied altitudes in UAV images result in more challenging pose variations and occlusions compared to conventional surveillance Re-ID datasets. Sample images from the AG-ReID.v2 dataset are shown in Figure 3. Before feature extraction, all the images were resized and normalized using the standard preprocessing pipeline provided by the CLIP model. This included resizing each image to 224×224 pixels, center cropping, and pixel normalization.

2.2. Person Region Extraction. The AG-ReID.v2 dataset provides pre-cropped images, where each image contains a person already extracted from a UAV-captured scene. Therefore, no additional person detection or cropping was required. Each image corresponds to an individual. These person-region images were passed directly to the CLIP model after preprocessing.

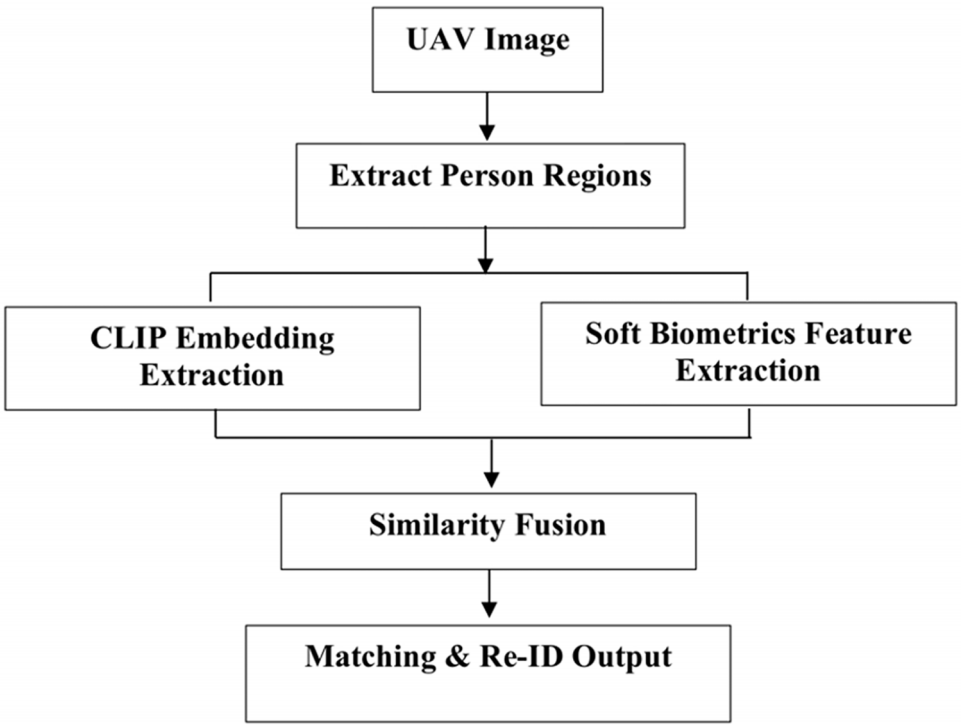


Figure 2. The proposed zero-shot UAV person Re-ID framework



Figure 3. Representative UAV-based images from the AG-ReID.v2 dataset

2.3. CLIP Embedding Extraction. Each cropped person image was passed through a CLIP image encoder to obtain a 512-dimensional visual feature vector. This vector serves as a compact representation of a person's appearance and captures high-level semantic characteristics from the image, as illustrated in Figure 4. These visual features were later used in similarity comparison and matching processes.

2.4. Soft Biometrics Feature Extraction. For each person's image, soft biometric features were extracted by comparing them with a set of predefined textual prompts using the CLIP model (Figure 5). These prompts describe interpretable traits such as gender, clothing, and accessories. The model assigned a score to each prompt, reflecting the closeness of the image to the described attribute. The result is a vector

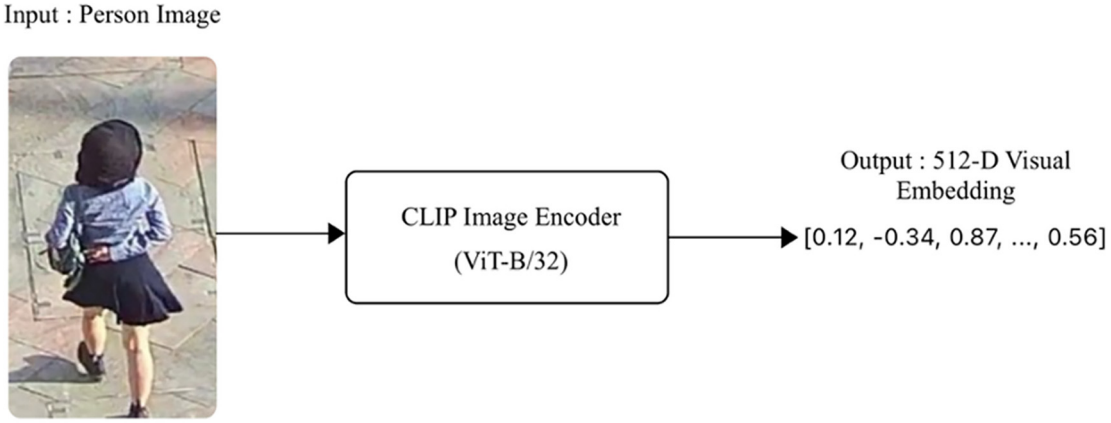


Figure 4. Image encoding using CLIP to generate an embedding vector



Figure 5. Extracted soft biometric attributes of the target

of scores representing the presence of different soft biometric traits.

2.5. Similarity Fusion. The proposed framework employs a fusion strategy that combines CLIP visual embeddings with soft biometric features to compute robust similarity scores for person Re-ID (Figure 6). The process begins by computing two similarity scores: visual similarity, which is obtained by applying the cosine similarity between the CLIP embeddings of the query image and gallery images; and biometric similarity, which is computed by measuring the cosine similarity between the soft biometric vectors of the query and gallery images (Figure 7). These similarity scores are then combined using a weighted summation, where visual similarity is assigned a weight of 0.8 and biometric similarity a weight of 0.2. The resulting value represents the final similarity score used for person matching.

2.6. Matching and Re-ID Output. To perform person re-identification, the system compares the fused similarity scores between the query image and each gallery image. This score combines the visual similarity (from CLIP embeddings) and soft biometric similarity. Based on these scores, all gallery entries are ranked, and the most similar candidates are selected as potential matches. More importantly, the system operates in a zero-shot setting, ensuring no overlap in identity between the query and gallery sets. An overview of the matching process is shown in Figure 8.

3. RESULTS AND DISCUSSION

The experiment was conducted using the AG-ReID.v2 dataset using a zero-shot protocol, that is, the model was applied directly without prior training. Implemented in Python with the PyTorch framework, the CLIP model (ViT-B/32) was employed to extract visual features from images without fine-tuning. Each image was processed using a CLIP encoder to generate a visual embedding vector. A set of predefined soft biometric prompts was then used to compute the similarity scores between each query and the gallery images. Both visual and biometric similarity scores were calculated and combined using a weighted summation, as defined in Equation (1), with greater emphasis placed on the visual component owing to the critical role of visual features in the re-identification task.

$$\begin{aligned} \textit{Final Score} = 0.8 \times \textit{Visual Similarity} + 0.2 \\ \times \textit{Biometric Similarity} \end{aligned} \tag{1}$$

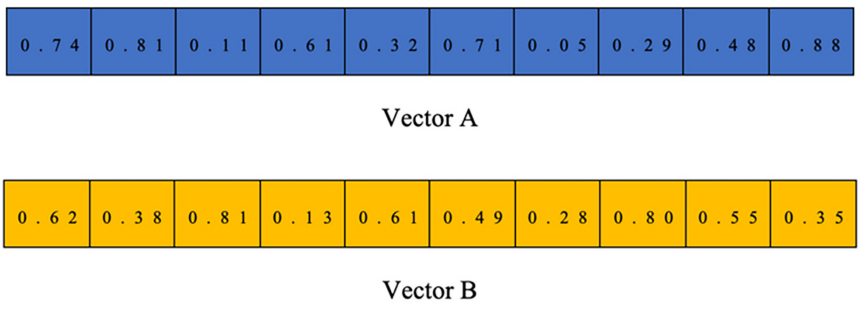


Figure 6. CLIP embeddings of vectors A and B

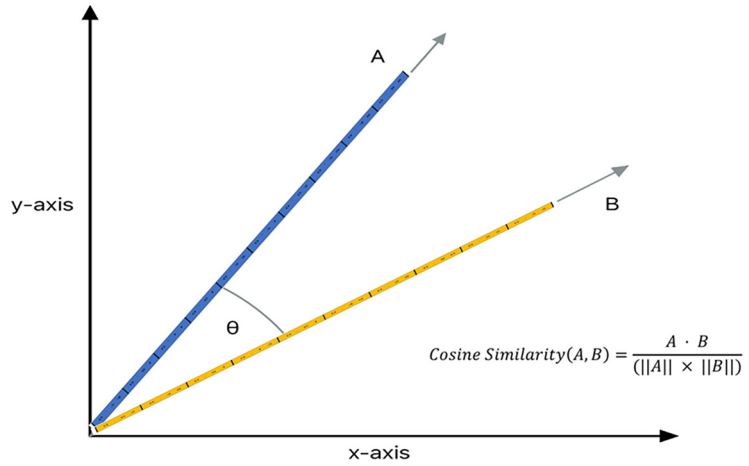


Figure 7. Measuring the similarity between two vectors, A and B, using the cosine similarity measure

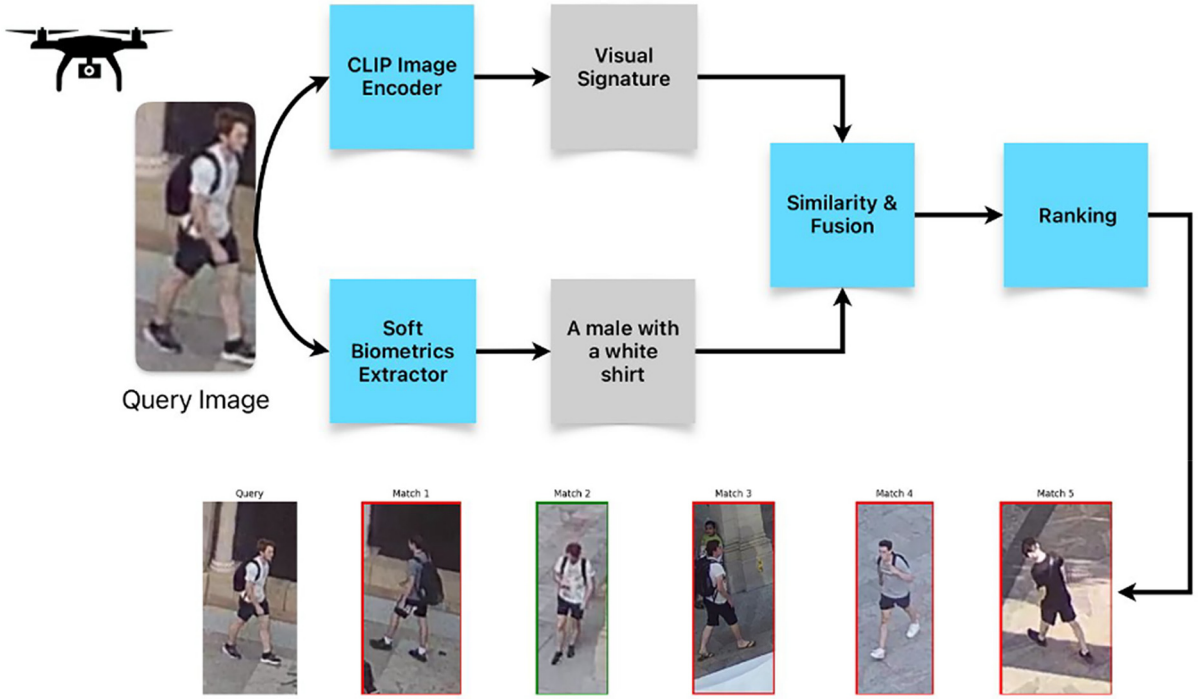


Figure 8. The process of matching a query image with the gallery data

3.1. Evaluation Measures. The evaluation metric used in this work was Rank-1 accuracy, which is a widely adopted measure for retrieval tasks. Rank-1 accuracy is defined as follows:

$$Rank - 1 = \frac{number\ of\ correct\ matches\ at\ rank - 1}{total\ number\ of\ queries} \quad (2)$$

3.2. Quantitative Results. To assess the effectiveness of the proposed CLIP-based person re-identification framework, its retrieval performance was evaluated using standard metrics, namely, Rank-1 accuracy. The system achieved a Rank-1 accuracy of 75.87% using the AG-ReID.v2 dataset. These results demonstrate the strong performance of the model in a zero-shot setting, where no gallery identities were encountered during setup or training. Notably, the integration of CLIP embeddings with soft biometric features enhances the matching robustness under challenging conditions such as viewpoint variations and partial occlusions. Compared with existing approaches in the literature (Table 2), the proposed method outperformed several trained models in terms of Rank-1 accuracy. It is important to highlight that while competing methods benefit from supervised training, the proposed approach operates in a zero-shot manner, underscoring its generalizability and practical applicability in scenarios with limited or no labeled data.

Table 2. *Performance comparison on UAV images from the AG-ReID.v2 dataset*

Method	Rank-1 Accuracy
Nguyen et al.[1]	72.75%
Xue et al.[7]	66.7%
Ours	75.87%

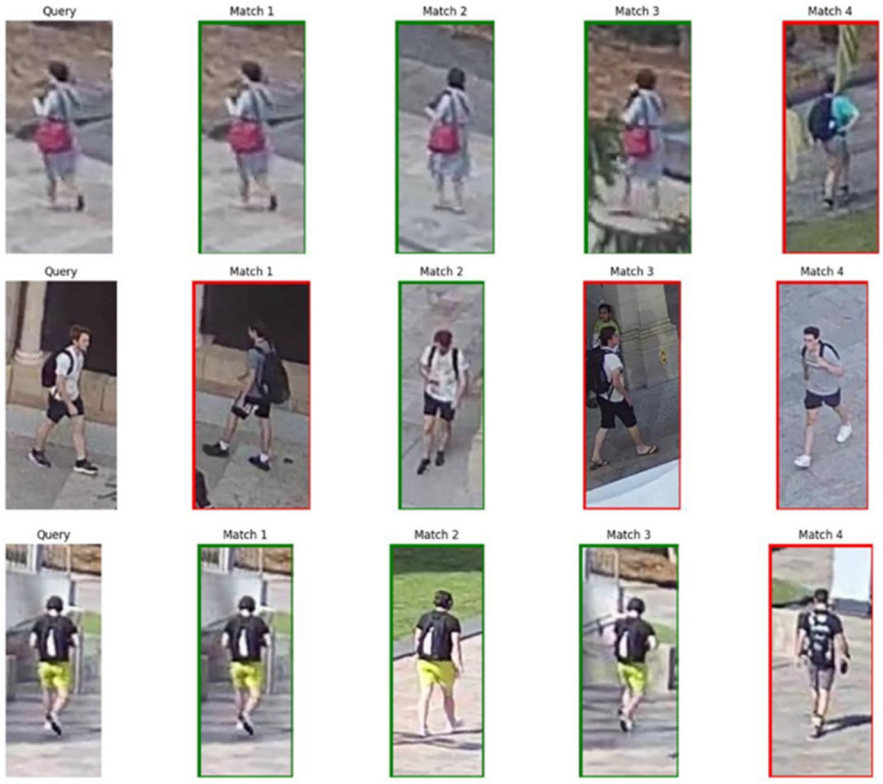
Table 3. *Ablation study on the effect of weighting coefficients*

Visual Similarity	Biometric Similarity	Rank-1
0.8	0.2	75.87%
0.5	0.5	75.76%

Ablation analysis was conducted to examine the effects of the weighting coefficients in Equation (1). The results are presented in Table 3. Increasing the weight assigned to the biometric similarity measure slightly reduces accuracy. This is because visual features are more important as they capture distinctive



(a) Successful Re-ID matches.



(b) Failure cases.

Figure 9. Illustrative results from the recognition task

characteristics, whereas supporting features, such as gender, are less informative. The computational cost depends on the size of the database and the hardware specifications. Generally, larger datasets and limited computational resources may lead to increased processing times.

3.3. Qualitative Results. To gain a deeper understanding of the system performance across various person re-identification scenarios, Figure 9(a) and 9(b) present visual examples of both successful and unsuccessful retrieval outcomes, illustrating how the system ranks and selects the most similar gallery images for a given query. Analysis of these examples reveals two predominant patterns: (a) successful matches, wherein the top five ranked images frequently correspond to the correct identity despite variations in camera viewpoints or background conditions, and (b) partially correct matches, where the initial ranks include correct identities followed by incorrect ones. The latter typically arises in cases involving individuals with visually similar appearances, such as comparable clothing or accessories, which can result in misidentification by the system. For instance, in the third row of Figure 9(b), a misidentified person carries a backpack that appears very similar to the query input. Nevertheless, in all these misclassified cases, the Rank-1 result was still correct, and most misidentified cases appear to occur after Rank-4.

4. CONCLUSIONS

This study explored a zero-shot person re-identification framework for UAV-captured imagery by leveraging CLIP-based visual embeddings combined with soft biometric features to enhance retrieval performance. Our approach combines semantic attributes with vision-language models, enabling efficient query-based searches without requiring identity-specific training data. The system demonstrated strong performance under aerial surveillance conditions, achieving a Rank-1 accuracy of 75.87%. The fusion of visual and soft biometric features enabled consistent ranking of correct identities within the top five matches, even in challenging cases with viewpoint variations or partial occlusions. These findings highlight the effectiveness of combining vision-language models with soft biometric features to improve the person re-identification performance in aerial surveillance applications. Future work will focus on refining attribute extraction and extending the framework to handle more challenging real-world scenarios.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Lama Al-Thunayyan – *Information & Computer Science Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0006-7882-0190
Email: s202275940@kfupm.edu.sa

Corresponding Author

Hussein Samma – *Research Mentor, Information & Computer Science Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia & SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0000-0002-3562-2788
Email: hussein.binsamma@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence.

REFERENCES

(1) Nguyen, H., Nguyen, K., Sridharan, S. & Fookes, C. AG-ReID.v2: Bridging Aerial and Ground Views for Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* 19, 2896–2908 (2024).

(2) Behera, N. K. S., Sa, P. K., Muhammad, K. & Bakshi, S. Large-Scale Person Re-Identification for Crowd Monitoring in Emergency. *IEEE Trans. Autom. Sci. Eng.* (2023).

(3) Gao, L. *et al.* Contrastive Pedestrian Attentive and Correlation Learning Network for Occluded Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* (2024).

(4) Yu, X., Dong, N., Zhu, L., Peng, H. & Tao, D. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Trans. Multimed.* (2025).

(5) Li, S. *et al.* Cocas+: Large-scale clothes-changing person re-identification with clothes templates. IEEE Trans. Circuits Syst. *Video Technol.* 33, 1839–1853 (2022).

(6) Wang, H. *et al.* Clothes-Changing Person Re-Identification Using Color and Texture Invariant Representation. *IEEE Trans. Intell. Transp. Syst.* (2024).

(7) Xue, C. *et al.* Vision transformer-based robust learning for cloth-changing person re-identification. *Appl. Soft Comput.* 111891 (2024).

(8) Nguyen, T. N. H. & Mai, T.-D. Enhancing Aerial-Ground Video Person Re-Identification via DFGS-Guided CLIP Sampling and Inference-Time Uncertainty-Aware Fusion. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 1591–1598* (2026).

(9) Hambarde, K. A. *et al.* VReID-XFD: Video-based Person Re-identification at Extreme Far Distance Challenge Results. ArXiv Prepr. ArXiv260101312 (2026).