ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Pedestrian Attribute Recognition Model for UAV Application: A Practical Approach

Kantida Parattanawong[1], Supaporn Erjongmanee[2], Eakarat Suwanagood[3] and Chaiwat Klampol[4]

ABSTRACT

Pedestrian Attribute Recognition (PAR) is an important component of intelligent surveillance systems. Ground-to-aerial cross-domain PAR and the effect of UAV flight conditions remain largely unexplored. This work investigates whether PAR models trained on ground-level CCTV datasets can be applied to UAV imagery and quantitatively analyzes the impact of UAV elevation angle and horizontal distance on attribute recognition performance. Five CNN models are trained on two CCTV datasets with different characteristics: UPAR, a large and diverse dataset, and TPAD, a homogeneous dataset, using a multi-label classification framework with positive class weighting to handle class imbalance. Cross-dataset evaluation on CCTV data leads to the selection of RegNet and ConvNeXt. For ground-to-aerial evaluation, selected models are evaluated on three UAV datasets: UAV-Human, AG-VPReID, and a self-collected UAV-PT1 dataset, where UPAR-trained models achieve a mean attribute accuracy of 62.23-65.48%, while TPAD-trained models perform worse. RegNet achieves comparable performance to ConvNeXt with significantly lower computational complexity, making it more suitable for UAV deployment. Attribute-level analysis shows that UpperBodyLength, LowerBodyLength, LowerBodyColor, and Backpack are more reliably recognized. Further analysis using UAV-PT1 shows increasing the horizontal distance from 25 m to 50 m reduces accuracy by 11.37–12.11%, and a high elevation angle of 50° causes a significant performance drop, providing an evaluation of ground-to-aerial PAR and the impact of UAV flight parameters on attribute recognition.

Article information:

Keywords: Pedestrian Attribute Recognition, Unmanned Aerial Vehicles, CCTV Datasets, UAV Images

Article history:
Received: March 1, 2025
Revised: November 16, 2025
Accepted: January 26, 2026
Published: January 31, 2026 (Online)

DOI: 10.37936/ecti-cit.2026201.260978

1. INTRODUCTION

In recent years, Unmanned Aerial Vehicles (UAVs) have demonstrated significant value across industries due to their cost-effectiveness and ability to access remote areas. Initially, UAVs were primarily used for intelligence, surveillance, target acquisition, and reconnaissance (ISTAR). However, advances in artificial intelligence (AI) have expanded their capabilities [1], transforming UAVs from simple aerial observation tools into sophisticated assets capable of supporting complex decision-making.

AI enables the rapid and accurate processing of large volumes of UAV data, transforming images into actionable intelligence that improves operational efficiency and decision making. Computer vision can analyze UAV images to detect patterns and extract valuable information for real-world applications such as border patrol, area surveillance, intelligence gathering, trespassing detection, and crowd monitoring. This AI integration into UAV systems highlights the growing importance of intelligent data processing in modern aerial operations.

One potential application of AI in UAV images is pedestrian attribute recognition (PAR). PAR supports tasks such as human identification, person re-identification, person search, tracking, and crowd analysis. However, UAV-specific datasets are scarce and often not publicly available due to challenges and

[1,2]The authors are with the Department of Computer Engineering, Kasetsart University, Bangkok, Thailand, E-mail: kantida.para@ku.th and fengspe@ku.ac.th
[3,4]The authors are with the Department of Aerospace Engineering, Kasetsart University, Bangkok, Thailand, E-mail: eakarat.s@ku.th and chaiwat.kl@ku.th
[2]Corresponding author: fengspe@ku.ac.th

restrictions in UAV data acquisition. As a result, annotated aerial image datasets are limited, prompting the need to explore alternative training data sources.

This work investigates the feasibility of using Closed-Circuit Television (CCTV) datasets to develop PAR models for UAV imagery. UAV images often suffer from low resolution, lighting conditions, ambiguous perspectives, and occlusions [2], making attribute recognition difficult. In contrast, CCTV images capture clearer views from lower altitudes, allowing more visible attributes. Since UAV data are rarely collected for surveillance tasks, PAR models often rely on publicly available CCTV datasets.

This study evaluates the performance of PAR models trained on CCTV data when applied to UAV imagery. It further examines model robustness under varying UAV-controlled factors, such as elevation angle and horizontal distance, as shown in Figure 1. Elevation angle introduces image perspective distortion [3], while increasing horizontal distance reduces image resolution. To our knowledge, no prior work has studied the impact of UAV-controlled factors on PAR performance in UAV imagery. Understanding these impacts can support developing robust PAR models across diverse UAV operational conditions.

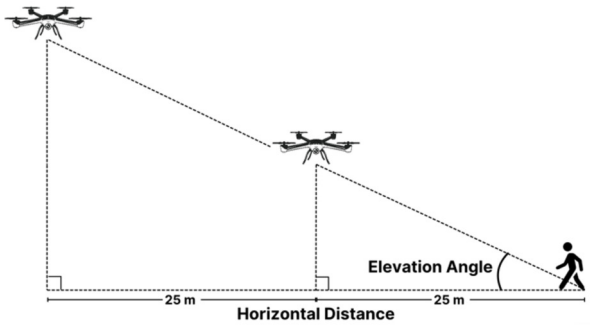


Fig.1: *Elevation angle and horizontal distance.*

Two CCTV datasets are used to develop the PAR models. The first, the UPAR [4] dataset, was collected across multiple locations and is considered a *diverse* dataset due to its wide range of pedestrian appearances and attributes. In contrast, the second dataset, TPAD [5], was collected at a single location and is therefore considered a *homogeneous* dataset, where it has limited variations of pedestrian attributes and appearances. The TPAD dataset was specifically constructed to support the development of PAR models tailored to Thai pedestrians and to represent local pedestrian attributes better. As reported in [5], PAR models trained on diverse datasets can exhibit non-local attribute biases, for example, over-predicting the long-sleeve attribute, which is uncommon in tropical regions like Thailand. Because TPAD was collected at a single site, it contains constrained pedestrian and attribute variations.

Five convolutional neural network (CNN) models are proposed for PAR-model construction. Given the limitations of power and computational resources for onboard UAV computing, two of these models are TinyML models capable of running directly on UAVs. Therefore, this work will consider the trade-off between computational efficiency and model performance across five CNN models trained on the CCTV datasets.

The PAR models trained on the CCTV dataset are validated using three UAV image datasets. Two of these, UAV-Human [6] and AG-VPReID [7], are publicly available datasets. The third dataset, UAV Perspective Test 1 (UAV-PT1), is self-collected for this study. UAV-PT1 was designed by varying two key UAV-controlled factors: elevation angle and horizontal distance. All three UAV datasets are used to evaluate the performance of CCTV-trained models, enabling the identification of the best- and worst-recognized pedestrian attributes and providing insights into which attributes are more reliably recognized in UAV imagery. For UAV-PT1, the effects of elevation angle and horizontal distance are further analyzed to assess their impact on PAR performance. Figure 2 summarizes the overall workflow of this study.

This work addresses the following question: Can CCTV-trained PAR models effectively recognize pedestrian attributes in UAV images captured at different elevation angles and horizontal distances? The findings will provide a foundation for using cost-effective CCTV datasets to develop more efficient PAR models for future UAV applications.

The paper is organized as follows. Section 2 reviews related works. Section 3 covers the methodology and evaluation. Section 4 presents the results and discussion. Finally, Section 5 provides the conclusion.

2. RELATED WORKS

Pedestrian image datasets have been used for tasks such as attribute recognition [6,8-10], pose estimation [6,10], and mostly for person tracking and re-identification [6,9-11]. The algorithms employed to tackle these problems include support vector machines [8,10], bag-of-words models [11], convolutional neural networks [9,10], and transformer networks [6]. This work focuses specifically on pedestrian attribute recognition.

2.1 Pedestrian attribute recognition: From Ground-level to Aerial Perspectives

Pedestrian attribute recognition has evolved significantly from traditional computer vision techniques to advanced deep learning methods. Early approaches, such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), primarily relied on hand-made features to identify attributes such as gender, clothing, and accessories. The stan-

dard PAR pipeline includes image pre-processing, feature extraction, and classification [2].

Recently, convolutional neural networks have become the dominant approach for feature extraction in PAR, enabling the models to learn hierarchical feature representations directly from raw image data, resulting in more accurate and robust feature extraction. Many CNN architectures have proven particularly effective for PAR. EfficientNet has been used as a baseline PAR model, achieving over 80% mA and recall across datasets such as PETA [8], Market-1501 [11], PA-100K [9], and RAP [10] without requiring additional model modifications [12]. DenseNet has been successfully applied to recognize pedestrian attributes from aerial views captured by UAVs [6]. More recently, ConvNeXt demonstrates promising results as a strong PAR model [13] and was employed as a baseline method for the UPAR Challenge 2024 [4].

Pedestrian images generally come from datasets of images [8] and videos [6,9-11]. Ground-level PAR has been studied using CCTV datasets captured at eye level or from slightly elevated positions [8-11], and these studies have significantly advanced the field. TPAD [5] demonstrates how locally collected attributes can impact the performance of PAR models. Sarfraz *et al.* [14] discuss the challenges posed by different pedestrian viewpoints, as attribute visibility varies significantly with viewing angles, directly impacting PAR system performance.

The transition from ground-level data to aerial images introduces unique challenges, including small object sizes, scale imbalance, and object rotation [15-17]. The UAV-Human dataset [6], constructed from UAV-based cameras, also highlights these challenges and shows that performance on UAV-based data is lower than on ground-level datasets.

2.2 Cross-Domain Adaptation for Surveillance Systems

Cross-domain adaptation is a major challenge in PAR. Models trained on one domain, such as CCTV images, often experience performance degradation when deployed in a different domain, such as UAV images. Differences in image resolution, viewpoint, background clutter, and occlusion mainly cause this degradation. Notably, unlike person re-identification (Re-ID), no existing work specifically studies PAR under a CCTV-trained/UAV-tested protocol.

The ground–aerial domain gap has been widely studied in person Re-ID. Wei *et al.* [18] show that Re-ID models trained on constrained datasets generalize poorly across domains due to variations in lighting, viewpoint, and camera conditions, leading to significant performance drops in cross-dataset evaluations. This issue is more pronounced in aerial imagery. Grigorev et al. [19] demonstrate that drone-captured images differ substantially from ground-based images in scale, resolution, and viewing angle, making direct transfer from CCTV-trained models ineffective. To address this problem, recent work introduces learning strategies for aerial imagery, such as meta-learning frameworks designed to improve generalization from CCTV to aerial domains [20]. However, these studies focus mainly on identity recognition and do not extend to attribute-level prediction tasks.

While cross-domain adaptation has been studied in PAR, existing work focuses on cross-dataset or cross-scenario generalization rather than the specific challenge of ground-to-aerial transfer. A recent PAR benchmark evaluates cross-domain partitioning by training and testing across different scene categories and reports an approximately 9% decrease in recall when evaluated on unseen domains [21]. Similarly, UPAR studies both leave-one-out and cross-dataset evaluation protocols, showing that the leave-one-out setting, which combines multiple training datasets, achieves better performance than cross-dataset evaluation due to the diversity of training data [13].

Overall, existing studies show that domain shifts significantly degrade performance in both Re-ID and PAR. While ground–aerial adaptation has been actively explored in person Re-ID, pedestrian attribute recognition has so far been studied only under ground-based or cross-dataset settings, and the problem of adapting PAR models from CCTV to UAV imagery remains largely unexplored. Our study focuses on the compatibility of applying models trained on a CCTV dataset to UAV imagery.

3. METHODOLOGY

The methodology describes the following: CCTV-trained PAR-model construction, UAV-dataset validation, and performance evaluation.

3.1 CCTV-trained PAR Models

CCTV cameras are widely used for surveillance, making them a cost-effective resource for PAR models, trained on collected CCTV datasets.

3.1.1 CCTV Training Datasets

The datasets for the PAR models in this study are obtain from two resources: publicly available UPAR and self-collected TPAD datasets.

A. UPAR Dataset

UPAR [4], released for the UPAR-Challenge 2024, focuses on real-world surveillance. UPAR combines three prior datasets: PETA [8], Market1501 [11], and PA100K [9], comprising a total of 131,076 images annotated with 40 attribute labels. Collected from multiple sources and locations, UPAR is considered a *diverse* dataset due to its wide range of attributes, pedestrian appearances, and scene conditions.

B. TPAD Dataset

To support the development of PAR models tailored to Thai pedestrians, the TPAD dataset [5] was

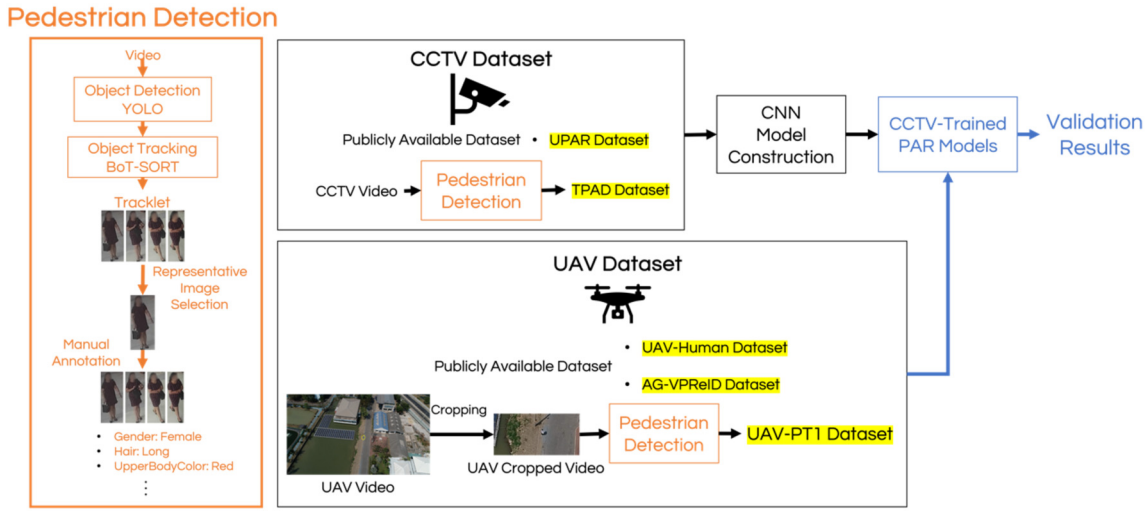


Fig.2: *Overall workflow.*

(a)Tracklet example 1 (b)Tracklet example 2

Fig.3: *Tracklet examples.*

Fig.4: *Tracklet-representative examples with various pedestrian attributes.*

constructed. Collected from a single location in Thailand, TPAD is regarded as a *homogeneous* dataset with limited attributes and pedestrian variability. The following describes the data acquisition and annotation procedures used to derive TPAD [5].

(i) *Image Processing:* Videos from multiple CCTV cameras in a single area were collected . A tracklet refers to a sequence of images across consecutive frames of the same individual. Figures 3(a) and 3(b) show two examples of a tracklet.

Pedestrians in each video frame are detected using the You-Only-Look-Once (YOLO) algorithm. YOLO is reportedly known for fast and accurate object detection and classification process [5]. Then, two multi-object tracking algorithms—ByteTrack [22] and BoT-SORT [23]—were evaluated. BoT-SORT enhances ByteTrack with a Kalman filter and camera motion compensation, resulting in higher tracking accuracy with only a minor loss in speed. Consequently, BoT-SORT was selected for tracklet extraction. Since BoT-SORT can occasionally assign the same ID to different individuals, all tracklets were manually reviewed. Any sequences with identity switches or images of occluded or partially visible pedestrians were discarded.

The image dimensions range from 28×69 pixels to 125× 311 pixels, with noticeable differences in lighting conditions. In total, 3,648 tracklets, comprising 12,185 pedestrian images, were obtained.

(ii) *Image Annotation:* With an average of 3.34 images per tracklet, attributes are assumed to be consistent within each tracklet. (Tracklet consistency was evaluated and reported in the Appendix.) Thus, a representative image is randomly selected for attribute annotation, and its attributes are applied to all images in the tracklet. Pedestrian images are manually annotated for demographics, upper- and lower-body appearance, and accessories. There are six binary and eight multi-class attributes, totaling 53 unique values, as shown in Table 1. Figure 4 presents representative tracklets with various attributes. For example, LowerBodyLength-Short is illustrated in Figures 4(a)-4(b), and LowerBodyLength-Long is in Figures 4(c)-4(h). LowerBodyType-Skirt&Dress is shown in Figure 4(d), while LowerBodyType-Trousers&Shorts are presented in the remaining figures.

[1] Permission to use the TPAD dataset is granted solely for scientific research, without public disclosure of the dataset. The analysis must be conducted anonymously and not at the level of personal data identification. Only statistical results are allowed to be reported. Note that Figures 3 and 4 were arranged to show examples of images similar to those in the TPAD dataset, but they are not part of the TPAD dataset. The actors in both figures have consented to the use of their images as examples in this paper.

Table 1: *TPAD attributes and their percentages.*

Group	Attribute Values (%)
Gender	Male (65.44%), Female (34.56%)
Age	Adult (86.66%), Senior (7.57%), Young (5.78%)
Hair	Short (76.51%), Long (20.22%), Bald(3.27%)
UpperBodyLength	Short (75.74%), Long (11.21%)
UpperBodyType	T-shirt (51.53%), Polo (7.73%), Tank top (7.37%), Shirt (5.79%), Blouse (3.36%), Jacket (2.40%), Dress (2.31%), Others (19.06%)
UpperBodyColor	Black (23.73%), White (21.81%), Blue (10.06%), Grey (7.65%), Green (6.86%), Red (5.15%), Brown (2.54%), Pink (2.48%), Orange (1.50%), Purple (1.50%), Yellow (1.43%), Others (14.63%)
LowerBodyLength	Short (56.20%), Long (43.80%)
LowerBodyType	Trousers&Shorts (93.35%), Skirt&Dress (6.65%)
LowerBodyColor	Black (44.35%), Blue(13.30%), Grey (7.04%), White(5.69%), Brown (3.39%), Green (2.84%), Pink (0.61%), Orange (0.54%), Red (0.32%), Purple(0.31%), Yellow (0.24%), Others (21.38%)
Footwear	Shoes (57.96%), Sandals (39.10%)
Glasses	Sun (5.88%), Normal (39.10%)
Hat	Yes (17.87%)
Bag	Yes (46.17%)
Backpack	Yes (10.79%)

3.1.2 Model Selection

Five CNN models are selected for evaluating image classification performance: ConvNeXt [24], EfficientNet [25], DenseNet [26], RegNet [27], and MobileNet [28]. The last two CNN models are TinyML models, generally feasible for deployment on resource- and power-constrained devices such as UAVs, mobile platforms, and embedded systems.

ConvNeXt modernizes traditional CNNs by incorporating design elements inspired by vision transformers, such as large convolution kernels and simplified stage-wise structures [24]. These changes improve global context modeling and representation stability. Nevertheless, early downsampling and large kernels can reduce spatial precision, limiting performance on fine-grained local attributes.

EfficientNet increases parameter efficiency by combining neural architecture search with compound scaling of depth, width, and resolution [25]. It also uses progressive resizing and adaptive regularization, leading to strong transfer learning performance. However, uniform scaling across the network may reduce sensitivity to small or localized attributes that require fine spatial detail.

DenseNet improves standard CNNs, such as ResNet-50, by connecting each layer to all subsequent layers, enhancing gradient flow, and enabling feature reuse [2,26]. This design supports faster convergence and parameter efficiency. However, strong feature reuse can reinforce dominant patterns from frequent classes, making DenseNet less effective for rare or subtle attributes under class imbalance.

RegNet uses a structured design strategy that defines network depth and width through simple, regular functions, enabling efficient scaling and stable learning across different computational budgets [27]. This regularity supports good generalization without increasing model complexity. However, the constrained design space may limit flexibility when modeling highly diverse or irregular visual patterns.

MobileNet is designed for low latency and minimal computational cost while maintaining reasonable accuracy [28]. It performs well in clean visual conditions with sufficient training samples. However, its lightweight architecture and limited channel capacity may reduce feature diversity, affecting recognition of complex or highly detailed attributes.

After the models are selected, their final classification layers are replaced with new trainable fully connected layers that output the requisite number of classes specifically for our datasets. This modification enables fine-tuning, allowing each model to learn more precise mappings for accurate classification. To prevent overfitting, dropout layers with a probability of 0.3 are incorporated into the models.

All models were trained for up to 100 epochs using mini-batch optimization. This upper bound was chosen to allow sufficient training for convergence across all model architectures while avoiding unnecessary computation, as preliminary experiments showed that performance typically plateaued well before reaching the maximum epoch limit. Early stopping was applied to prevent overfitting, and training was terminated if the validation loss did not improve for five consecutive epochs.

To address class imbalance in the multi-label classification setting, Binary Cross-Entropy with logits Loss was employed with class-wise positive weights. The positive weight for each class c was computed as $pos_{weight_c} = \frac{N}{N_c^+}$, where N denotes the total number of training samples and N_c^+ represents the number of positive samples for class c.

Hyperparameters were optimized using Optuna, selecting the batch size, initial learning rate, optimizer, weight decay, and learning rate scheduler for each architecture using model-specific search spaces. The optimal configurations are reported in the Appendix.

3.1.3 Experiment Setup

Before training the PAR models, the tracklet images undergo additional pre-processing. They are resized to a uniform 224×224 pixels for consistency. The dataset is then split into three subsets: 70% for training, 10% for use as a validation set, and 20% for

testing. To improve transfer learning, the images are normalized using mean and standard deviation values derived from a pre-trained ImageNet model. All random operations, including dataset splits and model initialization, were performed using a fixed random seed of 42 to ensure full reproducibility.

The experiments were conducted on two environments: (1) a workstation with an Intel i5-12400F CPU, an NVIDIA RTX 3060 GPU (12GB VRAM), and 16GB RAM, running Pop!_OS 22.04 LTS, and (2) a server with dual Intel Xeon Gold 6130 CPUs, NVIDIA Tesla V100-SXM2 GPU (32GB VRAM), running Ubuntu 20.04.4 LTS in a multi-user environment managed via SLURM, where computational resources were allocated per job (8 CPU cores, 32GB RAM, and 1 GPU for this work). Both systems run Python 3.10.12, with deep learning models implemented using PyTorch 1.12.1, TorchVision 0.16.1, and CUDA 12.1. For data acquisition, YOLO and BoT-SORT were used. Since the focus of this work is on models suitable for onboard operations on UAVs, the smallest pre-trained CNN models available in PyTorch were selected: ConvNeXt-Tiny [24], EfficientNet-V2-Small [25], DenseNet-121 [26], RegNetY-400MF [27], and MobileNet-V3-Small [28].

3.2 UAV Validation Datasets

CCTV-trained PAR models must be validated using actual UAV imagery. Three UAV datasets are used: two publicly available datasets, UAV-Human [6] and AG-VPReID [7], and one self-collected dataset, UAV Perspective Test 1 (UAV-PT1).

3.2.1 UAV-Human Dataset

UAV-Human is a publicly available UAV dataset for human behavior understanding, including tasks such as action recognition, person re-identification, and attribute recognition [6]. The dataset was collected over three months across 45 locations, at multiple altitudes, positions, and viewpoints, resulting in diverse UAV imagery. Images were captured at heights of 2-8 meters, leading to variations in resolution and occlusion, making the dataset representative of real-world UAV use. UAV-Human contains 22,263 images annotated across seven attribute groups, such as Gender, Hat, Backpack, Upper- and Lower-clothing color, and Style. To align its attributes with UPAR or TPAD, Upper- and Lower-clothing styles are converted to UpperBodyLength and LowerBodyType, respectively.

3.2.2 AG-VPReID Dataset

AG-VPReID dataset [7] is a large-scale, video-based dataset designed for aerial-ground person re-identification (Re-ID). It contains approximately 9.6 million frames, 32,321 tracklets, and 6,632 unique identities, captured over 20 days at nearby locations using drones at 15-120m altitude, CCTV ground cameras, and wearable cameras. Each identity is annotated with a comprehensive set of attributes, including gender, age, clothing style, and accessories. All identities matching and attribute annotation were manually verified by expert annotators, resulting in a total of 15 soft biometric attributes.

(a)10°, 25m (b)50°, 50m

Fig.5: *Original UAV image examples at different elevation angles and horizontal distances.*

(a)10°, 25m (b)50°, 50m

Fig.6: *Cropped UAV image examples at different elevation angles and horizontal distances.*

3.2.3 UAV-PT1 Dataset

The UAV Perspective Test 1 (UAV-PT1) dataset is added to assess the impact of two UAV-controlled factors: elevation angle and horizontal distance on the model performance. The elevation angle, determined by flying altitude, can cause perspective distortion and visual obstruction of the object. Horizontal distance affects image resolution and the level of detail captured. Greater horizontal distances may also amplify the effects of propulsive motor vibrations, resulting in motion blur and reduced image quality.

A. Attribute Selection

Due to limited UAV resources, it is not feasible to collect all 53 pedestrian attributes from TPAD. Initial experiments showed poor PAR performance on age and glasses, likely due to small facial features. Thus, these two attributes were excluded from the evaluation. For the remaining attributes, the most common attribute values were prioritized and included in our dataset. For example, black, white, and blue were selected for UpperBodyColor, while black and blue were selected for LowerBodyColor. In total, 27 attributes across 10 attribute groups were selected. During UAV image acquisition, actors wore clothing that matched the selected attributes.

B. Image Acquisition

UAV-PT1 images were captured using a DJI Phantom 4 Pro V2.0 quadrotor UAV, equipped with a 1-inch image sensor with a maximum resolution of 20

Table 2: *Model performance trained and tested on CCTV datasets.*

Model	GFLOPs	Trained and Tested on UPAR				Trained and Tested on TPAD			
		Params	Recall	F1	mA	Params	Recall	F1	mA
ConvNeXt	4.46	27.85M	*77.2*	*62.3*	*83.4*	27.86M	***92.5***	**92.6**	***94.7***
EfficientNet	*2.90*	20.22M	73.5	**66.2**	82.7	20.25M	**92.7**	**92.6**	**95.1**
DenseNet	*2.90*	*6.99M*	72.4	***63.0***	81.9	*7.01M*	91.8	90.9	94.2
RegNet	***0.40***	***3.90M***	**81.3**	60.0	**85.0**	***3.90M***	*92.2*	***91.4***	*94.5*
MobileNet	**0.06**	**1.56M**	***79.9***	58.4	***83.6***	**1.57M**	**92.7**	*91.3*	*94.5*

megapixels. Videos were recorded at a resolution of 2.7K (2720×1520 pixels).

Due to safety regulations, the maximum flying altitude was restricted to 60 meters, constraining the maximum elevation angle to 50°. In this study, the elevation angles were set to 10°, 20°, 30°, 40°, and 50°, and horizontal distances were 25 m and 50 m. Flying altitudes were pre-calculated based on these elevation angles and horizontal distances.

For acquisition, pedestrian locations were marked and horizontal distances were measured to enable accurate UAV positioning. Actors walked along predefined paths while the UAV took off to a pre-calculated flying altitude. The camera was adjusted to align the video frame with the walking path of each actor.

Since each actor walked toward the UAV for 5 meters, the specified horizontal distances corresponded to ranges of 20-25 m and 45-50 m. This results in a small deviation in elevation angles, ranging from 0.19-6.34° for 20-25 m and 0.19-3.96° for 45-50 m.

C. Image Processing

The recorded videos were processed. Figures 5(a)-5(b) show original UAV image examples captured at elevation angles of 10° and 50° and horizontal distances of 25 m and 50 m, with pedestrian location marked by yellow circles. As the elevation angle or horizontal distance increases, the pedestrian appears smaller and the image resolution decreases. At high elevation angles, strong winds can also make UAV control difficult, causing slight shift in the pedestrian's walking path.

As flying distance increases, image resolution degrades, making pedestrians very small. Thus, the UAV images were cropped to remove irrelevant areas, as shown in Figure 6. Cropped image dimensions range from approximately 490×410 to 690×510 pixels (4.83-8.47% of the original image) at 25 m, and 280×210 to 590×230 pixels (1.41-3.26%) at 50 m. High elevation angles make pedestrians appear more distant and upright, while large horizontal distances reduce apparent size.

Unlike the TPAD CCTV images, pedestrian detection in UAV-PT1 images could not be processed with YOLOv8. Therefore, the larger YOLOv11 model was used, resulting in a UAV-PT1 dataset of 1,728 images annotated with 27 attributes.

The validation process involves applying the CCTV-trained PAR models to the three UAV datasets to evaluate their performance.

3.3 Evaluation Metrics

Given that TP, FP, and FN respectively are numbers of correct positive, incorrect positive, and incorrect negative predictions, metrics used to evaluate the PAR-model performances in this work are:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

$$mA = \frac{1}{2A} \sum_{i=1}^{A} \left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i} \right) \tag{4}$$

where TP_i, TN_i, P_i, and N_i represent the true positives, the true negatives, the total positives, and the total negatives for attribute i. Generally, mean accuracy (mA) evaluates the overall classification performance across multiple classes.

To compute the attribute recognition accuracy, the following metric is used.

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{5}$$

where $N_{correct}$ is the number of correctly classified samples, and N_{total} is the total number of samples.

The detection quality of the YOLO model is evaluated using mAP_{50-95}, which considers classification accuracy and localization precision. It is computed as:

$$mAP_{50-95} = \frac{1}{|T|} \sum_{t \in T} AP_t \tag{6}$$

where $AP_t = \int_0^1 p_t(r)dr, p_t(r)$ is the precision-recall curve at IoU threshold t, and T is the set of IoU threshold values {0.50,0.55,...,0.95}.

4. RESULTS AND DISCUSSION

This section presents the performance of CCTV-trained models evaluated on UAV validation datasets.

Table 3: *Cross- and same-domain performance evaluated on UAV datasets.*

Train set	Validation set	RegNet			ConvNeXt		
		Recall	F1	mA	Recall	F1	mA
UPAR	UAV-Human	**51.76**	**38.50**	**66.47**	**51.36**	**40.25**	**67.79**
TPAD		28.83	24.53	56.71	33.52	31.52	59.41
UAV-Human		29.20	27.65	56.18	29.56	28.26	56.26
AG-VPReID		50.88	47.81	51.55	39.99	39.61	49.22
UPAR	UAV-Human	44.13	35.77	54.08	38.69	32.36	52.34
TPAD		31.22	25.98	50.14	36.36	27.98	51.33
UAV-Human		56.14	39.77	50.91	50.50	38.10	51.10
AG-VPReID		**99.93**	**99.85**	**98.50**	**99.93**	**99.59**	**99.85**
UPAR	UAV-PT1	**68.63**	**62.74**	**69.23**	**59.52**	**56.09**	**65.38**
TPAD		43.67	39.82	56.63	54.01	49.97	61.58
UAV-Human		45.21	35.91	53.88	*48.70*	36.96	56.05
AG-VPReID		51.78	38.81	54.06	41.67	32.79	50.52

4.1 CCTV-Trained Model Results

Performance of the constructed CCTV-trained models is evaluated, with a detailed discussion.

4.1.1 CCTV-Trained Model Performance

The performance of five PAR models, trained and tested on the UPAR and TPAD CCTV datasets, is shown in Table 2. The top three performances are highlighted using bold, bold with italics, and italics, respectively. Overall, performance across models is relatively similar. For the UPAR dataset, the highest performing models are RegNet, MobileNet, and ConvNeXt. For the TPAD dataset, the top-performing models are EfficientNet and ConvNeXt. Performance on TPAD, a homogeneous dataset, is higher than on UPAR, a diverse dataset, due to the lower variability in the TPAD test data. A detailed discussion explaining the observed model performance is presented next.

4.1.2 CCTV-Trained Model Discussion

Figures 7 and 8 present attribute-wise recall comparisons across models. Green and red cells denote relatively high and low recall, respectively, while grey cells represent similar recall across all models. Attributes not shown in the figures exhibit comparable recall values across all models (i.e., all grey cells in the same row), indicating no model recognizes these attributes distinctively.

On the UPAR dataset, many green-highlighted attributes are dominated by RegNet and MobileNet, particularly for accessory-related, color-based, and rare demographic attributes. Their strong performance can be attributed to their architectural simplicity and regularity. RegNet adopts a uniform and well-structured architecture, enabling stable feature learning across layers without excessive feature reuse or overly large receptive fields. This design allows RegNet to generalize more effectively under class imbalance, leading to strong recall for attributes, such as Age-Old, Age-Young, sunglasses, and several rare

	ConvNeXt	EfficientNet	Densenet	RegNet	MobileNet
Accessory Bag	67.8	76.2	68.1	72.2	73.8
Age Old	66.1	67.7	66.1	75.3	71.7
Age Young	55.3	54.7	50.0	67.4	64.7
Glasses Sun	32.4	28.4	30.8	41.9	32.6
Glasses Normal	80.4	75.1	69.9	84.9	84.7
HairLength Bald	44.2	39.7	40.2	56.2	48.2
LowerBodyColor-Purple	50.0	50.0	50.0	50.0	58.3
LowerBodyColor-Other	38.1	31.9	21.3	52.1	50.7
LowerBodyColor-Green	58.6	44.2	44.2	63.1	53.4
LowerBodyColor-Red	78.0	66.1	72.9	93.2	93.2
LowerBodyColor-Yellow	78.8	57.6	54.5	63.6	57.6
LowerBodyColor-White	80.4	78.0	74.1	83.8	86.5
LowerBodyColor-Pink	82.9	73.4	76.4	86.3	80.6
LowerBodyColor-Blue	84.1	80.4	77.0	88.4	86.7
LowerBodyColor-Brown	89.9	85.2	83.2	92.5	89.6
LowerBodyColor-Grey	83.9	81.1	72.9	85.9	87.6
LowerBodyType Skirt&dress	74.0	79.0	77.2	78.8	80.7
UpperBodyColor-Orange	82.4	65.9	68.2	85.9	85.9
UpperBodyColor-Pink	74.2	64.0	69.1	84.8	85.4
UpperBodyColor-Grey	79.7	78.4	72.0	84.7	84.9
UpperBodyColor-Blue	80.5	74.4	74.3	83.0	85.5
UpperBodyColor-Purple	81.2	80.0	76.9	88.1	85.0
UpperBodyColor-Green	82.6	78.3	81.9	90.3	86.9
UpperBodyColor-Other	57.0	56.8	52.3	73.8	71.3

Fig.7: *Comparison of attribute-wise recall across UPAR-trained models.*

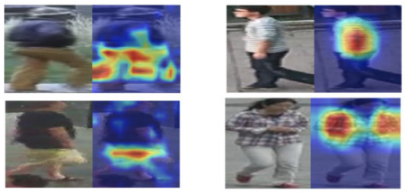
	ConvNeXt	EfficientNet	Densenet	RegNet	MobileNet
Accessory Bag	96.8	95.9	96.0	93.7	96.8
Glasses Normal	86.5	86.5	82.5	87.3	87.3
Glasses Sun	80.8	82.4	80.8	86.4	83.2
LowerBodyColor-Purple	80.0	80.0	80.0	80.0	100.0
LowerBodyColor-Pink	73.1	73.1	69.2	73.1	73.1
LowerBodyColor-Green	92.2	96.1	93.5	93.5	92.2
UpperBodyColor-Orange	84.4	86.7	88.9	86.7	86.7
UpperBodyColor-Yellow	90.3	90.3	93.5	90.3	90.3
UpperBodyColor-Purple	89.5	100.0	94.7	92.1	94.7
UpperBodyType Shirt	84.1	82.5	80.2	84.1	84.9
UpperBodyType Polo	94.7	95.2	95.2	94.7	92.3
UpperBodyType Dress	95.0	90.0	83.3	88.3	88.3

Fig.8: *Comparison of attribute-wise recall across TPAD-trained models.*

colors. Similarly, MobileNet employs depthwise separable convolutions that preserve local spatial information while avoiding excessive feature mixing, making it suitable for recognizing small or fine-grained objects. The TPAD results further support this observation, with both RegNet and MobileNet consistently achieving competitive or best recall for many fine-grained attributes, even when overall recall is high across all models.

ConvNeXt shows stable but not dominant performance across attributes on UPAR. This behavior aligns with its architectural design, which uses large 7×7 convolution kernels and early downsampling. Such a structure favors global context modeling and color consistency, leading to reliable performance on large-scale attributes. However, the gradual reduction in feature map resolution can cause small or localized features to become blurred or disappear in deeper layers. As a result, ConvNeXt does not consistently achieve the best recall for fine-grained attributes. Similar performance results are observed on the TPAD dataset.

EfficientNet shows balanced performance on UPAR but demonstrates stronger results on TPAD. Its compound scaling strategy jointly optimizes network depth, width, and resolution, enabling efficient and robust feature learning. On UPAR, EfficientNet performs competitively on common color and clothing attributes but provides limited gains for rare or fine-grained attributes. In contrast, on TPAD, EfficientNet achieves a higher recall for several color-based and accessory-related attributes, indicating improved generalization across a wider range of attributes.



(a)ConvNeXt (LowerBodyColor-Yellow) (b)RegNet (UpperBodyColor-Other)

Fig.9: *Grad-CAM visualizations for selected attributes.*

DenseNet performs relatively poorly on UPAR for many low-frequency attributes, especially rare colors and Age-Young groups. Although its dense connections promote feature reuse, this mechanism tends to reinforce dominant patterns from frequent attributes, increasing bias toward common classes. This behavior is also evident on TPAD, where DenseNet occasionally underperforms on less frequent attributes despite strong performance on common categories. These results suggest that feature reuse alone is insufficient for learning rare or subtle attributes.

Overall, results from both UPAR and TPAD indicate that architectures preserving spatial detail and avoiding overly dominant feature reuse are better suited for challenging pedestrian attributes. ConvNeXt is effective for global and color-consistent attributes, while DenseNet is more sensitive to data imbalance due to its feature reuse strategy. In contrast, MobileNet and RegNet demonstrate robust performance for rare, small-scale, and fine-grained attributes across both datasets. Grad-CAM visualizations further support this observation. ConvNeXt exhibits broader activation patterns as presented in Figure 9(a), while RegNet focuses on more localized and discriminative regions as shown in Figure 9(b). These findings confirm that architectural inductive bias, rather than model complexity, plays a key role in pedestrian attribute recognition performance.

Table 4: *Attribute recognition performance.*

Attribute Group	UAV-Human	AG-VPReID	UAV-PT1
Gender	56.85	37.78	73.44
Hat	44.56	85.65	89.93
Hair		69.21	54.98
UpperBodyLength	74.23		75.06
UpperBodyColor	87.04		37.96
LowerBodyLength		83.88	62.04
LowerBodyColor	78.26		63.95
LowerBodyType	40.71		53.99
Bag		52.23	38.66
Backpack	69.98	64.12	72.69
Average	64.52	65.48	62.23

Considering both UPAR- and TPAD-trained models, ConvNeXt, RegNet, and MobileNet, show relatively strong overall performance. However, while UPAR-trained MobileNet achieved performance comparable to RegNet, it recognized fewer attributes. Consequently, RegNet and ConvNeXt are selected as representative models for evaluation on UAV datasets.

4.2 UAV Data Validation Performance

This section evaluates whether the two CNN models—RegNet and ConvNeXt—trained on CCTV datasets can recognize pedestrian attributes on three UAV datasets: UAV-Human, AG-VPReID, and UAV-PT1. For comparison, models trained on UAV-Human and AG-VPReID are also included. Due to its small size, UAV-PT1 is used only for evaluation.

4.2.1 Cross- and Same-Domain Performance

Table 3 summarizes the validation performance of CCTV-trained and UAV-trained PAR models on three UAV datasets.

When comparing CCTV-trained models from UPAR and TPAD on the same UAV dataset, UPAR-trained models consistently outperform the TPAD-trained models. Even on UAV-PT1, which shares local attributes with TPAD, the UPAR-trained model

achieves better results. This finding indicates that models trained on a diverse dataset generally generalize better than those trained on a homogeneous dataset, highlighting the importance of dataset diversity for CCTV-trained PAR models.

Comparing UPAR-trained with UAV-trained models from UAV-Human and AG-VPReID, UPAR-trained models achieve better performance on UAV-Human and UAV-PT1. This suggests that UAV datasets may be less suitable for training PAR models due to lower image quality.

However, the AG-VPReID-trained model performed relatively better on its own validation set. This strong in-dataset performance can be attributed to the large size of AG-VPReID and its collection from nearby locations, effectively representing a single-location homogeneous dataset with low attribute variability. The combination of a large number of images and homogeneous attributes leads to very high in-dataset performance.

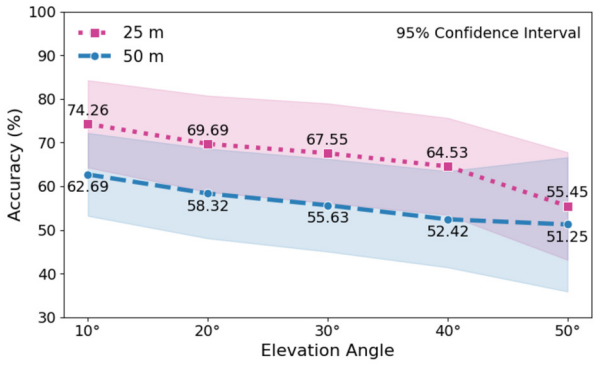


Fig.10: *Model performance on UAV factors.*

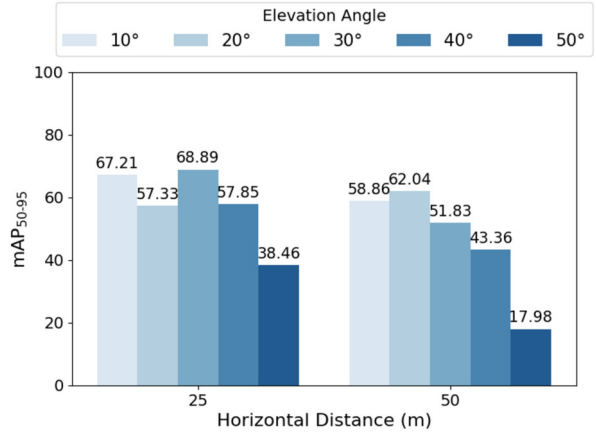


Fig.11: *YOLO performance on UAV factors.*

Overall, these results indicate that for developing PAR models applicable to diverse UAV imagery, training on a large and diverse CCTV dataset is generally preferable. Exceptions occur in single-location scenarios: when a large UAV dataset is available from a single location, it may be more effective to train the PAR model directly on this UAV dataset rather than on CCTV images.

When comparing RegNet and ConvNeXt, their overall performance is relatively similar, with ConvNeXt performing slightly better. Nonetheless, due to its large GFLOPS and total parameter count, ConvNeXt is less feasible for onboard deployment. Therefore, RegNet is selected as the representative model for further UAV data validation in this work.

4.2.2 Attribute Recognition Performance

Table 4 presents the attribute recognition performance of the UPAR-trained RegNet model evaluated on three UAV datasets. Cells without reported values indicate that such a dataset does not annotate those attributes. The attribute recognition accuracy ranges from 62.23% to 65.48% on average. This performance is largely attributable to the inherent challenges of UAV imagery, including varying flight altitudes, reduced spatial resolution, and the small visual size of pedestrians.

Despite these challenges, several attributes demonstrate consistent and relatively strong recognition performance across the evaluated datasets, notably UpperBodyLength (sleeve length; average 74.64%), LowerBodyLength (average 72.96%), LowerBodyColor (average 71.10%), and Backpack (average 68.93%). In contrast, Hat and UpperBodyColor achieve high accuracy in certain datasets but exhibit limited generalization across all datasets. Other attributes, including Gender, Hair, LowerBodyType, and Bag, demonstrate lower recognition performance.

These results suggest that Hat, as well as torso-related attributes, such as UpperBodyLength, LowerBodyLength, LowerBodyColor, and Backpack, are reliably recognized attributes in UAV images. These attributes serve as more discriminative and robust features for pedestrian recognition in UAV datasets.

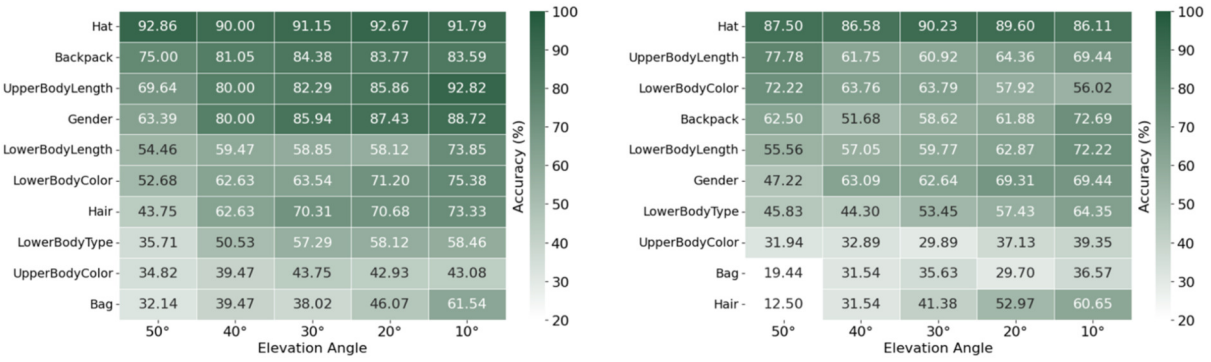
4.3 Impacts of UAV-controlled Factors

The impacts of UAV-controlled factors–horizontal distance and elevation angle–on PAR performance are reported. UAV-PT1 is used in this section, as it contains data collected across varying horizontal distances and elevation angles.

4.3.1 Impacts on Model Performance

Figure 10 shows the 95% confidence intervals of accuracy for UPAR-trained models evaluated on the UAV-PT1 datasets across elevation angles from 10° to 50° at two horizontal distances (25 m and 50 m).

Increasing the horizontal distance from 25 m to 50 m reduces accuracy by 11.37-12.11% at 10-40°, and by 4.20% at 50°. At 25 m, each 10° increase in elevation angle causes accuracy drops of 4.57%, 2.13%, 3.038%, and 9.08%. In contrast, at 50 m, the corresponding decreases are 4.37%, 2.68%, 3.22%, and 1.17%.



(a)25 m (b)50 m

Fig.12: *Attributes performance on UAV factors.*

These results show that accuracy slightly declines at low-to-moderate elevation angles (10-40°) for both distances. At 50°, the accuracy drop is substantial at 25 m (9.08%) but declines only marginally at 50 m (1.17%). At longer distances, pedestrians occupy small regions. Further increases in elevation angle have a limited impact on recognition accuracy. The two-way ANOVA results indicate that horizontal distance significantly affects recognition accuracy, whereas elevation angles between 10° and 40° show no significant impact. In contrast, an elevation angle of 50° shows a significant effect on accuracy.

Figure 11 presents the detection performance of the YOLO model measured by mAP_{50-95}. For both 25 m and 50 m, performance declines at 10-40° and drops sharply at 50°. Therefore, detection performance steadily declines with increasing elevation angle and degrades more sharply at longer horizontal distances.

4.3.2 Impacts on Attributes

The impact of UAV-controlled factors on pedestrian attribute recognition is analyzed by evaluating UPAR-trained models on the UAV-PT1 dataset.

Figures 12(a) and 12(b) present heatmaps of attribute recognition performance under varying UAV-controlled conditions. The numbers in heatmaps indicate recognition accuracy at different elevation angles. Attributes are ranked in descending order based on accuracy at the highest elevation angle (50°), which is placed in the leftmost column.

Overall, attribute recognition accuracy decreases as elevation angle increases, regardless of horizontal distance. Attributes such as Gender, Hair, UpperBodyLength, LowerBodyType, Bag, and Backpack show moderate to substantial performance degradation as the elevation angle increases. In contrast, Hat and UpperBodyColor are minimally affected by changes in elevation angle, indicating stronger robustness to viewpoint variation. Interestingly, under the extreme condition of 50° elevation angle and 50 m horizontal distance, UpperBodyLength, LowerBodyColor, and Backpack exhibit contrasting behavior, showing improved accuracy. These results indicate that the effect of elevation angle is not uniform across attributes but varies depending on the specific attribute. Horizontal distance further influences performance, and its interaction with elevation angle ultimately determines overall recognition accuracy.

For fixed elevation angles between 10-40°, increasing horizontal distance from 25-50m leads to a pronounced performance drop for Hair. Backpack, UpperBodyLength, Gender, and Bag show moderate degradation, whereas Hat, UpperBodyColor, LowerBodyLength, and LowerBodyType are slightly affected by increased distance. At a high elevation angle of 50°, increasing horizontal distance leads to noticeable accuracy improvements for UpperBodyLength, LowerBodyColor, and LowerBodyType, while most other attributes continue to decline.

Overall, attribute recognition accuracy varies across different combinations of UAV-controlled factors. The two-way ANOVA results are used to analyze the attribute-specific impacts as follows:

- Both horizontal distance and elevation angle significantly affect Gender, Hair, and Bag.
- Horizontal distance has a significant impact on Hat, UpperBodyColor, and UpperBodyLength.
- Elevation angle significantly affects Backpack, LowerBodyLength, and LowerBodyType.
- Under extreme conditions (high elevation angle combined with long horizontal distance), both factors jointly influence Backpack, UpperBodyLength, LowerBodyColor, and LowerBodyType.

Comparing these results with Table 4 further demonstrates that attribute recognition performance is strongly influenced by UAV-controlled factors. These findings highlight the importance of accounting for UAV-controlled factors when developing pedestrian attribute recognition models for UAV applications and provide insights into which attributes remain reliable under challenging operational conditions.

5. CONCLUSION

This work studied pedestrian attribute recognition in a ground-to-aerial cross-domain setting, addressing the limited analysis of applying CCTV-trained PAR models to UAV imagery and the lack of understanding of how UAV flight conditions affect recognition performance. Five CNN-based PAR models were first trained on CCTV datasets using a multi-label classification framework with positive class weighting to handle class imbalance. Cross-dataset evaluation showed that models trained on the diverse UPAR dataset generalize better than those trained on the homogeneous TPAD dataset. Based on cross-validation results, only RegNet and ConvNeXt were selected for evaluation on UAV datasets. Among them, RegNet was selected for detailed UAV analysis due to its lower computational complexity. The selected model was tested on three UAV datasets, including the self-collected UAV-PT1 dataset, where the effects of elevation angle and horizontal distance were analyzed. The results show that higher elevation angles and longer distances degrade attribute recognition performance, with varying effects across attributes. This work has several limitations, including potential noise from tracklet-level labels, the limited size of the UAV-PT1 dataset, and detection bottlenecks caused by extremely small pedestrians in UAV images. Future work will focus on collecting larger UAV datasets and exploring domain adaptation methods to enhance PAR robustness in real UAV applications.

ACKNOWLEDGEMENT

This work was financially supported by the Master's Degree Study Support Program (Plan A) of the Faculty of Engineering, Kasetsart University. The authors also acknowledge the support of NontriAI, Office of Computer Services, Kasetsart University, Bangkok, Thailand, for providing computational resources, technical support, and infrastructure for model development and experimentation. The authors would like to thank all contributors involved in UAV operations and data collection for their valuable assistance.

AUTHOR CONTRIBUTIONS

Conceptualization, K.P., S.E., E.S., and C.K.; methodology, K.P., S.E., E.S., and C.K.; software, K.P.; validation, K.P. and S.E.; formal analysis, K.P. and S.E.; investigation, K.P.; data curation, K.P.; writing—original draft preparation, K.P.; writing—review and editing, S.E.; supervision, S.E., E.S., and C.K. All authors have read and agreed to the published version of the manuscript.

References

[1] A. K. Sachdev, "Artificial intelligence and military aviation," in *Artificial Intelligence, Ethics and the Future of Warfare, Routledge India*, pp. 91–107, 2024.

[2] X. Wang *et al.*, "Pedestrian attribute recognition: A survey," *Pattern Recognition*, vol. 121, p. 108220, 2022.

[3] A. G. Perera, A. Al-Naji, Y. W. Law and J. Chahl, "Human detection and motion analysis from a quadrotor UAV," in *IOP Conference Series: Materials Science and Engineering*, vol. 405, no. 1, p. 012003, 2018.

[4] M. Cormier *et al.*, "UPAR Challenge 2024: Pedestrian Attribute Recognition and Attribute-Based Person Retrieval - Dataset, Design, and Results," *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, pp. 359-367, 2024.

[5] K. Parattanawong, S. Erjongmanee, E. Suwanagood and C. Klumpol, "Benchmarking pedestrian attribute recognition systems for UAVs using locally collected dataset: A case study in Thailand," in *28th International Computer Science and Engineering Conference*, pp. 1–6, 2024.

[6] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang and Z. Li, "UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 16261-16270, 2021.

[7] H. Nguyen, K. Nguyen, A. Pemasiri, F. Liu, S. Sridharan and C. Fookes, "AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 1241-1251, 2025.

[8] Y. Deng, P. Luo, C. C. Loy and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 789–792, 2014.

[9] X. Liu *et al.*, "HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 350-359, 2017.

[10] D. Li, Z. Zhang, X. Chen and K. Huang, "A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios," in *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575-1590, April 2019.

[11] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-identification: A Benchmark," *2015 IEEE International Con-*

ference on Computer Vision (ICCV), Santiago, Chile, pp. 1116-1124, 2015.

[12] D. Weng, Z. Tan, L. Fang and G. Guo, "Exploring attribute localization and correlation for pedestrian attribute recognition," *Neurocomputing*, vol. 531, pp. 140–150, 2023.

[13] A. Specker, M. Cormier and J. Beyerer, "UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 981-990, 2023.

[14] M. S. Sarfraz, A. Schumann, Y. Wang and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," *arXiv preprint* arXiv:1707.06089, 2017.

[15] V. Pandey, K. Anand, A. Kalra, A. Gupta, P. P. Roy and B.-G. Kim, "Enhancing object detection in aerial images," *Mathematical Biosciences and Engineering*, vol. 19, no. 8, pp. 7920–7932, 2022.

[16] G. Tang, J. Ni, Y. Zhao, Y. Gu and W. Cao, "A survey of object detection for UAVs based on deep learning," *Mathematical Biosciences and Engineering*, vol. 16, no. 1, p. 149, 2023.

[17] J. Leng *et al.*, "Recent advances for aerial object detection: A survey," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–36, 2024.

[18] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 79-88, 2018.

[19] A. Grigorev, Z. Tian, S. Rho, J. Xiong, S. Liu and F. Jiang, "Deep person re-identification in UAV images," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, no. 54, 2019.

[20] L. Xu, H. Peng, L. Wang and D. Xia, "Meta-transfer learning for person re-identification in aerial imagery," *Computer Supported Cooperative Work and Social Computing*, pp. 634–644, 2022.

[21] J. Jin, X. Wang, Q. Zhu, H. Wang and C. Li, "Pedestrian attribute recognition: A new benchmark dataset and a large language model augmented framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 4138–4146, 2025.

[22] Y. Zhang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," *Computer Vision – ECCV 2022*, pp. 1–21, 2022.

[23] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations for multi-pedestrian tracking," *arXiv preprint* arXiv:2206.14651, 2022.

[24] Z. Liu, H. Mao, C. -Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A ConvNet for the 2020s," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 11966-11976, 2022.

[25] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261-2269, 2017.

[27] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He and P. Dollár, "Designing Network Design Spaces," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 10425-10433, 2020.

[28] A. Howard *et al.*, "Searching for MobileNetV3," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 1314-1324, 2019.

Table 5: *Optimal hyperparameters for each dataset and backbone architecture.*

Dataset	Model	LR	BS	Opt	WD	Sched	Milestone	γ	Step	Min LR
UPAR	ConvNeXt-Tiny	2.32×10^{-5}	16	Adam	2.97×10^{-4}	Cosine	–	–	–	–
UPAR	EfficientNet-B0	4.36×10^{-4}	32	AdamW	8.41×10^{-4}	Cosine	–	–	–	–
UPAR	DenseNet-121	6.67×10^{-5}	32	AdamW	4.65×10^{-5}	MultiStepLR	[16, 68]	0.498	–	–
UPAR	RegNetY-400MF	4.04×10^{-4}	32	Adam	6.21×10^{-5}	Cosine	–	–	–	–
UPAR	MobileNetV3-Small	1.00×10^{-4}	32	Adam	1.98×10^{-5}	Cosine	–	–	–	–
TPAD	ConvNeXt-Tiny	5.00×10^{-5}	32	AdamW	1.00×10^{-2}	Cosine	–	–	–	1.0×10^{-6}
TPAD	EfficientNet-B0	7.07×10^{-4}	64	AdamW	1.63×10^{-6}	StepLR	–	0.195	41	–
TPAD	DenseNet-121	1.81×10^{-4}	32	AdamW	7.72×10^{-6}	MultiStepLR	[17, 48]	0.423	–	–
TPAD	RegNetY-400MF	3.23×10^{-4}	16	Adam	1.32×10^{-5}	Cosine	–	–	–	–
TPAD	MobileNetV3-Small	1.19×10^{-3}	32	Adam	3.93×10^{-6}	Cosine	–	–	–	–

APPENDIX

A. TRACKLET ANNOTATION CONSISTENCY

To validate attribute consistency within tracklets, 10% of tracklets were randomly sampled and fully annotated at the image level. For each image, the disagreement rate was computed as $(FP + FN)/N_{attr}$, where N_{attr} denotes the total number of evaluated attributes per image. Tracklet-level disagreement was obtained by averaging across images. The overall disagreement rate was 2.55%, indicating that a single representative image provides a reliable approximation for tracklet-wise attribute annotation.

B. HYPERPARAMETER SETTINGS

Table 5 summarizes the optimal hyperparameter configurations used for training each backbone architecture on the UPAR and TPAD datasets.

These settings were selected based on validation performance and were fixed for all reported experiments.

Kantida Parattanawong received the B.Eng. degree in Computer Engineering from Kasetsart University, Thailand, in 2024. She is currently pursuing the M.Eng. degree in Computer Engineering at the Department of Computer Engineering, Kasetsart University. Her research interests include computer vision, machine learning, person re-identification, pedestrian attribute recognition, and UAV-based surveillance systems.

Supaporn Erjongmanee is an Assistant Professor in the Department of Computer Engineering at Kasetsart University, Bangkok, Thailand. She received her B.S. (Honors) degree in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, USA, in 2001, and her M.S. and Ph.D. degrees in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, USA, in 2003 and 2011, respectively. Her research interests include applied machine learning and artificial intelligence, with an emphasis on UAV applications.

Eakarat Suwanagood is a Research Officer with the Department of Aerospace Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand. He received the B.Eng. degree in Aerospace Engineering through a dual-degree program from Kasetsart University, Thailand, and the Royal Melbourne Institute of Technology (RMIT University), Melbourne, Australia, in 2012. His research interests include small unmanned aerial vehicle (UAV) aerodynamics, unmanned aerial systems (UAS) design, and the application of machine learning and artificial intelligence for enhancing the performance and autonomy of future UAV systems.

Chaiwat Klampol works as University's administrator, researcher, and lecturer at Kasetsart University. His research works at the Department of Aerospace Engineering include Aircraft and Spacecraft Structural Dynamics, Artificial Intelligent Model Development for Aerial Detection, Tracking and Automatic Control of Unmanned Aerial Vehicles. He received his B.Eng. in Mechanical Engineering from Kasetsart University in 1997, and his M.S. and Ph.D. in Aerospace Engineering from the University of Colorado in 1999 and 2004 respectively. His research interests include inverse modeling, system identification, and adaptive control.