An Efficient Pedestrian Attribute Recognition System under Challenging Conditions

Ha X. Nguyen[1,3,*], Dong N. Hoang[3], Tuan A. Tran[2,3], and Tuan M. Dang[3,4,5]
[1]*Research Group Intelligent Robots, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam*
[2]*School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam*
[3]*CMC Applied Technology Institute, CMC Corporation, 11 Duy Tan, Hanoi, Vietnam*
[4]*CMC University, CMC Corporation, 11 Duy Tan, Hanoi, Vietnam*
[5]*Posts and Telecommunication Institute of Technology, KM 10 Nguyen Trai, Ha Dong, Hanoi, Vietnam*
[*]*Corresponding author: Ha X. Nguyen (ha.nguyenxuan@hust.edu.vn)*

Abstract. In this work, an efficient pedestrian attribute recognition system is introduced. The system is based on a novel processing pipeline that combines the best-performing attribute extraction model with an efficient attribute filtering algorithm using keypoints of human pose. The attribute extraction models are developed based on several state-of-the-art deep networks via transfer learning techniques, including ResNet50, Swin-transformer, and ConvNeXt. Pre-trained models of these networks are fine-tuned using the Ensemble Pedestrian Attribute Recognition (EPAR) dataset. Several optimization techniques, including the advanced optimizer Adam with Decoupled Weight Decay Regularization (AdamW), Random Erasing (RE), and weighted loss functions, are adopted to solve issues of data unbalancing or challenging conditions like partial and occluded bodies. Experimental evaluations are performed via EPAR that contains 26 993 images of 1477 person IDs, most of which are in challenging conditions. The results show that the ConvNeXt-v2-B outperforms other networks; mean accuracy (mA) reaches 85.57%, and other indices are also the highest. The addition of AdamW or RE can improve accuracy by 1-2%. The use of new loss functions can solve the issue of data unbalancing, in which the accuracy of data-less attributes improves by a maximum of 14% in the best case. Significantly, when the attribute filtering algorithm is applied, the results are dramatically improved, and mA reaches an excellent value of 94.85%. Utilizing the state-of-the-art attribute extraction model with optimization techniques on the large-scale and diverse dataset and attribute filtering has shown a good approach and thus has a high potential for practical applications.

Key words: pedestrian attribute recognition, Deep Learning, vision transformer, security surveillance.

1. Introduction

Pedestrian Attribute Recognition (PAR) is an area of computer vision that tries to assess and comprehend the characteristics of people shown in still images and moving videos. The purpose of the PAR system is to automatically extract and classify features such as clothing style, things that a pedestrian is carrying, and physical factors such as age, gender, and ethnicity from photos and videos of pedestrians. The information PAR

gleans may be used in various applications, including image retrieval, human-computer interaction, and video surveillance [2].

The variable looks of humans, the existence of occlusions, and the constantly shifting lighting conditions all contribute to the difficulty of measuring PAR. There have been many advances made in PAR [1,6,8,11,15,17,19,34,35,38]. However, there are still a lot of obstacles to overcome, such as dealing with complex and complicated circumstances, increasing recognition accuracy, and lowering the computational processing requirement of PAR systems. In recent years, deep learning strategies have seen widespread use in PAR due to their encouraging results in learning complicated and hierarchical representations of human characteristics. Deep neural networks are incredibly effective in PAR and are now being used by several cutting-edge computer systems [31].

Most of the existing works regarding PAR models have been adopted via off-the-shelf pre-trained deep networks as their backbone network architecture. The pre-trained deep networks are often developed based on large-scale datasets such as ImageNet [4]. Most techniques exploit the ResNet50 [9] as the backbone. Recently, some novel deep networks have been proposed, such as Swin-transformer [21] and ConvNeXt [22, 33]. Although these networks are based on the structure of modern vision transformers with many advancements, they need to be adapted to match the unique characteristics of PAR systems. Thus, novel deep networks should be further developed. For example, in [3], Cheng *et al.* proposed a new model architecture to achieve the best accuracy on the RAP and PA-100K datasets. However, scaling the introduced model up with extensive backbones depends a lot on the size of the embedding words of the textual module. Therefore, in the development of novel backbones there is still much work to do.

Most of the published works use well-known datasets like PA100K [20], PETA [5], RAPv2 [16], MSMT17 [32], and Market1501 [18] as benchmarks for both training and testing phase. Recently, a unified dataset named UPAR to allow generalization experiments for 40 attributes across four PAR datasets PA100K [20], PETA [5], RAPv2 [16], and Market1501 [18] was proposed [28]. These studies concentrated on one particular setting, such as an indoor or outdoor environment, and use relatively large-scale datasets. Nevertheless, there are many restrictions surrounding the generation of real-world surveillance datasets. Challenging conditions including multi-view, unbalanced data distribution, occlusion, low resolution, and poor or varying illumination, are the main issues influencing the final recognized performance. Therefore, augmentation techniques for large-scale and diverse datasets and training optimization facing these issues are current research trends.

Besides the dataset strategies, there have been several adaptions for the transfer learning techniques of the state-of-the-art deep networks to overcome challenging conditions. The issues of data unbalancing can be solved by using suitable loss functions. Many new loss functions, such as (Weighted) Cross Entropy Loss, Contrastive Loss,

Center Loss, Triplet Loss, and Focal Loss, have been suggested for the optimization of deep neural networks. Recently, several novel advanced loss functions for PAR have been reported [14, 29, 36]. Also, the Random Erasing (RE) technique [37] is introduced to overcome the issue of the occluded or partial body. The use of novel optimizers, like Adam [12], or Adam with Decoupled Weight Decay Regularization (AdamW) [23], can also improve the model's accuracy. It has been known that the use of advanced techniques for the transfer learning process can improve quite a small amount of recognition accuracy.

Our primary goal in this research is to develop an efficient pedestrian attributes recognition system in challenging conditions scenarios. The proposed system will have not only a remarkably high accuracy but also the robustness against challenging recognition conditions of practical applications.

The main contributions of this work are as follows:

- Proposing a novel processing pipeline combining feature extractions models with keypoints-based attributes filtering for recognizing pedestrians' attributes from challenging condition scenarios.
- Generating an ensemble person attribute recognition dataset based on several state-of-the-art datasets, considering the diverse and challenging scenarios.
- Applying the transfer learning technique selected from among various deep network architectures, including ResNet50, Swin-transformer, and ConvNeXt, to obtain the best-performing model.
- Tuning the best-performing model by adapting training optimization techniques like AdamW, Random Erasing, and advanced loss functions to overcome issues of data unbalancing and partial or occluded body.
- Systematically evaluating experimental results on a challenging dataset and producing system design instructions for practical applications.

In the following section, detailed descriptions of the proposed system are presented.

2. Proposed system

The overall processing pipeline of the proposed system is shown in Fig. 1. Unlike in the most of the published works [7, 27], which have concentrated only on models for feature extraction to achieve higher accuracy on typical datasets, our system is adapted by using the keypoint concept as an additional information channel for the prediction. There are three main modules, each responsible for a specific task. Image frames from camera streams are pushed into the human body, pose detection and tracking module. As a result, the bounding box and corresponding pose of detected persons are achieved and consequently tracked. Images and poses of tracked persons are then pushed into the attributes recognition module for attributes prediction and the color recognition module for body color prediction. The use of keypoints of the pose has the advantage

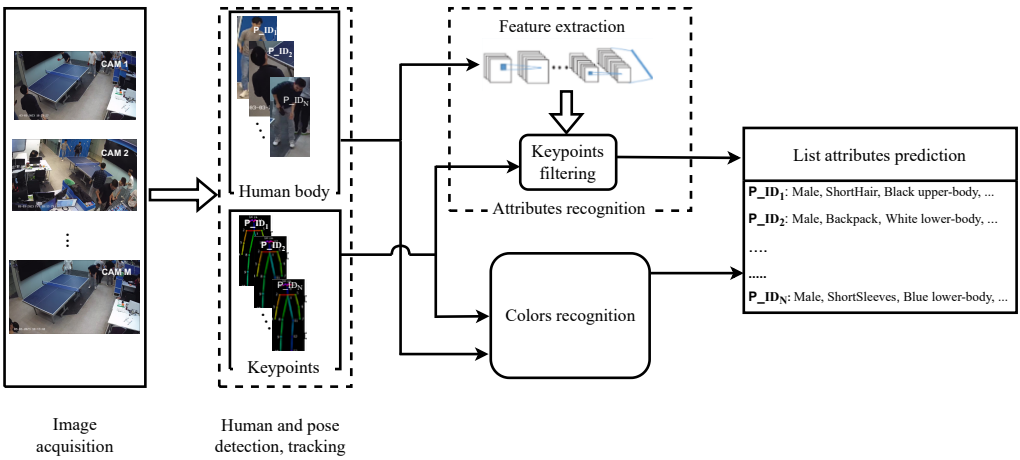

Fig. 1. Overview of our processing pipeline.

of improving the accuracy and robustness of the system in challenging conditions such as occluded or partial detecting bodies and viewpoint-variant bodies. In the following, details of each module are described.

2.1. Ensemble Pedestrian Attribute Recognition dataset

The ensemble dataset (EPAR) was created by collecting and processing image data from the four publicly accessible datasets, including PA100K [20], MSMT17 [32], RAPv2 [16], and Market1501 [18]. We have labeled the data in the test set of the datasets mentioned above using the Pseudo-Label approach [13] in addition to the data in the training set. We have set the classification threshold to 0.9 and utilized the data labeling tool to accurately align the labels using pre-trained weights previously trained on the RAPv2 dataset. We then have got the EPAR dataset by joining them all. The ensemble of these datasets has the benefit that the EPAR has a massive variation in situations, races, and qualities. Images of EPAR are very diverse and cover many challenging conditions, including significant variations in pose, lighting conditions, background, and occlusions. For practical applications in security surveillance scenarios, 13 typical attributes of the human body outfits are chosen to label, including “Male”, “Female”, “Adult”, “Children”, “Long-Hair”, “ShortHair”, “Hat”, “Long-Sleeves”, “Short-Sleeves”, “Trouser-Jeans”, “Skirt”, “Short”, and “Backpack”. Since images of PA100k, MSMT17, and Market1510 are not assigned to person IDs, the re-identification method from [26] was used for the person

Tab. 1. Statistics portion of each dataset contributing to the EPAR.

Dataset from	No. IDs	No. Imgs	Properties
RAPv2 [16]	2589	30 315	annotated with viewpoints, occlusions, and body parts information by multiple cameras in real-world environments
PA100K [20]	3556	36 920	images are blurry due to the relatively low resolution and the positive ratio of each binary attribute is low
Market1501 [18]	700	12 668	images in this dataset exhibit significant variations in pose, lighting conditions, background, and clothing
MSMT17 [32]	600	10 101	images are captured in morning, noon, and afternoon in campus
EPAR (**ours**)	7445	90 004	all of above properties

ID register. The EPAR dataset contains 90 004 images of 7445 person IDs, with annotations for 13 binary attributes. Tab. 1 lists the portion and properties of each referring dataset contributing to the EPAR. EPAR is divided into 60%, 20%, and 20% for the test, train, and validation set. Fig. 2 shows the frequency of appearance of each attribute in all 90 004 images of EPAR. It is clear that the dataset EPAR has a big issue of data unbalancing, where some attributes, for example, "Children" and "Short" have a low frequency of appearance, making EPAR more challenging.

2.2. Human body and pose detection and tracking

Video frames from cameras are processed using Yolov7-Pose [24, 30]. The outputs are the bounding boxes and keypoints of the pose of detected persons. The use of Yolov7-Pose has the benefit that this model can simultaneously detect persons and estimate their poses, allowing us to save computational hardware resources. Also, this model outperforms others in accuracy and robustness in challenging detection scenarios. Since the images of a detected person can appear in several video frames of different cameras, the re-identification method [26] is used to match these images together and assign them to a tracking ID. Consequently, images and poses of each tracking ID with different viewpoints from different cameras are continued processing in the following steps. Since attributes of a person can be predicted from viewpoint-different images, the loss of information regarding the viewpoint can be maximally reduced.

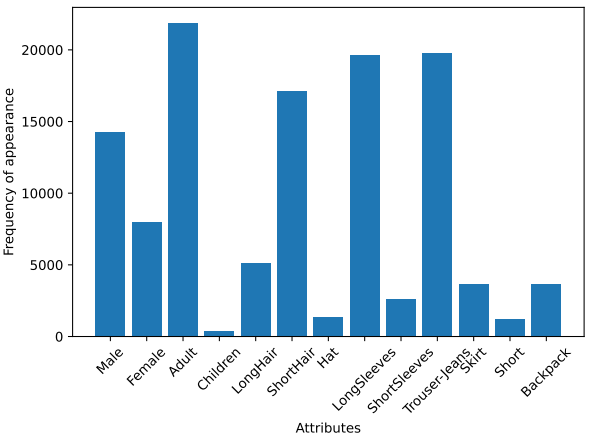

Fig. 2. Statistics of frequency of appearance of each attribute in EPAR.

2.3. Human attributes extraction models

Images of tracked persons are continued processing to extract corresponding attributes via an attributes extraction model. There are several approaches using convolution neural networks [9], visual transformer [21], or a hybrid of those like ConvNeXt [22, 33]. The ConvNeXt is considered to be state-of-the-art in accuracy. Thus, in this work, transfer learning techniques are applied to the ConvNeXt-v2-B [33] to achieve the attributes extraction model. Also, to evaluate the performance of the ConvNeXt-v2-B, other backbones including ResNet50 [9], Swin-Transformer [21], and ConvNeXt-v1-B [22], are fine-tuned. Consequently, the performance and accuracy of these models are compared and analyzed. All fine-tuning processes are performed using the EPAR dataset.

For the best performance of the developed model, the transfer learning technique is adapted to make several improvements. First, instead of using the conventional optimizer like Adam [12], the newly introduced one, namely the AdamW [23], was used. AdamW brings an improvement by incorporating weight decay into the Adam algorithm. In the original Adam algorithm, weight decay is usually applied by adding a term to the gradient of the parameters. However, this can lead to the loss of certain important properties of the Adam algorithm. In AdamW, weight decay is computed differently by directly applying it to the weights instead of modifying the gradient. This helps preserve the invariance properties of gradient scale and enhances the stability of the training process. Second, besides typical data augmentation techniques of the baseline, like random flipping, random gray-scale, and cropping, the Random Erasing [37] was

additionally used. This helped the training process to stay balanced and to ensure the diversity of the training data. The third improvement was regarding the loss function. A suitable loss function is expected to solve data unbalancing, as shown in Fig. 2. The loss function of the baseline training method is based on the cross-entropy formulated as

$$\mathcal{L} = -\sum_{j=1}^{M} w_j \left(y_{ij} \log\left(p_{ij}\right) + \left(1 - y_{ij}\right) \left(1 - \log\left(p_{ij}\right)\right)\right) , \tag{1}$$

where w_j is the attribute weight function of j^{th} attribute; p_{ij} is the output of the classifier layer. The three most popular weight function methods called $L1$ [14], $L2$ [29], and $L3$ [36], were experimented and evaluated. The description of each weight function is detailed in Tab. 2

2.4. Keypoints-Based Attributes Filtering

In fact, many recognition situations exist where images of the human body are occluded or partial, caused by obstacles or view-point variances. If only the attributes extraction model is used, the model always produces predictions for all attributes, even in cases where some of these attributes are occluded or do not appear in images. This leads to false predictions and thus reduces the system's accuracy. This issue is solved by using the keypoints as additional information for the attributes prediction. The pose's keypoints and confidence scores will let us know which body part in the images is occluded or

Tab. 2. Types of weight function used in the transfer learning process.

Method	Weight function
$L1$ by Li *et. al.* [14]	$w_j = \begin{cases} e^{1-r_j} & \text{when } y_{ij} = 1\,, \\ e^{r_j} & \text{when } y_{ij} = 0\,; \end{cases}$
$L2$ by Tan *et. al.* [29]	$w_j = \begin{cases} \sqrt{\dfrac{1}{2r_j}} & \text{when } y_{ij} = 1\,, \\ \sqrt{\dfrac{1}{2(1-r_j)}} & \text{when } y_{ij} = 0\,; \end{cases}$
$L3$ by Zhang *et. al.* [36]	$w_j = \begin{cases} \dfrac{\frac{1}{r_j^{\alpha}}}{\frac{1}{r_j^{\alpha}} + \frac{1}{(1-r_j)^{\alpha}}} & \text{when } y_{ij} = 1\,, \\ \dfrac{\frac{1}{(1-r_j)^{\alpha}}}{\frac{1}{r_j^{\alpha}} + \frac{1}{(1-r_j)^{\alpha}}} & \text{when } y_{ij} = 0\,; \end{cases}$
where α is a hyper-parameter to adjust the weight between positive ratio and negative ratio, and r_j is the positive sample ratio of j-th attribute in the training set	

partial. The attributes belonging to the occluded or partial parts will not be predicted. This will help the system avoid false predictions. The pose inferred by the Yolov7-Pose has 17 keypoints arranged to three parts of the human body, including the head, upper body, and lower body. An algorithm for attributes filtering was developed as illustrated in Algorithm 1. For each part of the human body, if more than half of the number of keypoints in part have a confidence score smaller than a threshold, this part will be concluded to be occluded or partial.

2.5. Keypoints-based body's color recognition

The keypoints were also used for predicting the color of the upper and lower body parts. An algorithm was proposed for this prediction. For the upper-body parts, a pair of keypoints, including the "right shoulder" and "left hip", is used to calculate the center of the upper-body part, which is the middle point of this pair of keypoints. Similarly, the "left hip" and "left ankle" are used for the lower-body part. The color of each interesting body part is calculated from the average of the color value of all pixels in a square with dimensions of 20×20 at the center of the part. The color of each pixel is computed using functions of OpenCV library [25] in the RGB mode.

3. Results and discussion

An evaluation dataset was created to evaluate the proposed system's accuracy and robustness. The dataset covered 17 677 images of the test dataset of EPAR and 9252 augmented images. The augmented images in the evaluating dataset were to add challenging conditions for the evaluation. As illustrated in Fig. 3, challenging situations in real-life applications, including partial or occluded bodies, images with low resolution, or inadequate lighting conditions, were considered. The evaluation dataset covered 26 933

Algorithm 1: Attributes Filtering.

Input: logit keypoints p, keypoints confidence score k, threshold T.
Output: filter logit p^*.
parts = $k[a:b]$
count = 0
for *part in parts* **do**
  if *part* $< T$ **then**
    count = count + 1
end
if *count > (number of keypoints in parts)/2* **then**
  $p[c:d] = 0$
$p^* = p$

Fig. 3. Illustration of images in challenging conditions in the evaluating dataset.

Tab. 3. Evaluation results of the proposed system.

Method	Backbone	mA	Acc	Prec	Recall	$F1$
Baseline [10]	ResNet50 [9]	82.05 ± 0.14	77.97 ± 0.23	80.78 ± 0.14	93.75 ± 0.28	86.78 ± 0.20
	Swin-S [21]	84.61 ± 0.19	79.52 ± 0.92	81.59 ± 0.69	95.33 ± 0.43	87.92 ± 0.59
	ConvNeXt-v1-B [22]	85.04 ± 0.39	79.85 ± 0.57	81.95 ± 0.47	95.41 ± 0.33	88.08 ± 0.35
	ConvNeXt-v2-B [33]	85.57 ± 0.28	80.04 ± 0.54	82.19 ± 0.38	95.88 ± 0.19	88.08 ± 0.29
+ AdamW [23]	ConvNeXt-v2-B	85.67 ± 0.02	81.43 ± 0.07	83.09 ± 0.07	96.06 ± 0.06	89.11 ± 0.07
+ Random Erasing [37]	ConvNeXt-v2-B	85.76 ± 0.05	81.63 ± 0.06	83.99 ± 0.08	$\mathbf{96.15 \pm 0.13}$	89.15 ± 0.05
+ Keypoints (**ours**)	ConvNeXt-v2-B	$\mathbf{94.28 \pm 0.01}$	$\mathbf{91.57 \pm 0.07}$	$\mathbf{94.43 \pm 0.07}$	93.65 ± 0.07	$\mathbf{93.61 \pm 0.06}$

images of 1477 person IDs, including 30% partial, 15% occlusion, 23% outdoor, 7.5% indoor, 10% gray-scale, and 15% normal-quality images. For metrics, four instance-level metrics and one attribute-level (label-level) measure were used to evaluate the model's performance based on the literature [20]. Accuracy (Acc), Precision (Prec), Recall (Recall), and $F1$ are used as metrics at the instance level. Mean accuracy (mA) is used as an attribute-level statistic since it focuses on the recognition accuracy of a particular attribute. The instance-level metric is used when we want to evaluate the model's overall performance in terms of its ability to predict entire instances. This approach is particularly relevant in regression tasks or when the primary concern is the quality of instance-level predictions rather than individual label predictions. The attribute-based criteria are used when we want to assess the model's performance with a focus on individual labels or classes. These criteria are helpful for understanding the model's precision, recall, and accuracy for each category, which can be especially useful when dealing with imbalanced datasets.

The evaluation results of the system are shown in Tab. 3. In the second-row cluster, results with the baseline method with backbone ResNet50, Swin-S, ConvNeXt-v1-B, and ConvNeXt-v2-B are compared. The third-row cluster shows the results of our improvements based on the ConvNeXt-v2-B with several adaptions using: i) AdamW; ii) AdamW and RE, and iii) AdamW and RE and keypoints, respectively. Each result

in the table is an average of ten repeated tests accompanied by the deviation. It is seen that, with the baseline method, if different backbones are used, the mA and other parameters are slightly changed. The ConvNext-v2-B, in which the mA reaches 85.57%, outperforms the other networks. When other optimizations, like AdamW or AdamW and RE, are used, the accuracy is also increased, but the increase is tiny at 0.1-0.2%. With our proposed method using keypoints as the filter, the mA increases significantly from 85.76% to 94.28%. Similarly, the Acc, Prec, and $F1$ also increase. Only the Recall decreases from 96.15% to 93.65%. This issue can be explained by that the pose detection model used in this work is just a pre-trained version of Yolov7-Pose without any modification. Thus, this still has failures in some detection scenarios. Consequently, keypoints-based filtering can eliminate some attributes which appear in images leading to a reduction of the True Positive Rate and the Recall.

Influences of the weight functions on the system's accuracy and data unbalancing are also evaluated. Tab. 4 shows the evaluation results of the system on different weight functions listed in Tab. 2. It can be seen that the weight function method L3 [36] slightly outperforms others in mA and Prec. This improvement is explained that, as stated by Zhang *et al.* [36], the $L3$ weight function (with $\alpha = 1$) helps re-weight for attributes with low frequency of appearance. As a result, $L3$ increases the True Negative Rate and thus reduces the incorrect recognition probability for attributes. That means the mA is improved. As Zhang *et al.* [36] mentioned in their paper, when setting $\alpha = 1$, $L3$ can help attributes-balanced re-weight, which increases mA. With $\alpha = 0$, the loss function transforms into conventional CE loss to help instance-balanced re-weight, which increases instance-level indexes such as Accuracy, Precision, Recall, and $F1$. It is inferred that the weight function can help improve the system's accuracy, especially in unbalanced data cases. However, the improvement is not much in the 0.1% (94.85% vs. 94.04%).

Detailed analysis of the True Positive Rate results for attributes with a low appearance frequency in the EPAR is shown in Fig. 4. The results of considering weight functions are presented. It is seen that the use of new weight functions improves the True Positive Rate significantly. For example, with the "Children" attribute, when the $L3$ is used, the True Positive Rate increases to 14%, compared to the conventional loss function. Similar trends for other considered attributes are also confirmed. These results have proven the efficient role of weight functions in solving the issue of data unbalancing.

Tab. 5 shows evaluated results of the computational requirement of the system. This work utilized a server with NVIDIA Tesla V100 32GB GPUs equipped with PyTorch 1.7.1 and CUDA 10.1 for the training and evaluation processes. It can be seen that although the ConvNeXt and Swin-S back-bone have more parameters as well as computing complexity requirements, the processing time is three times less than that of ResNet50 (210ms vs. 70ms). This can be clarified that the ConvNeXt or Swin-S have a new inverted bottlenecks design which helps reduce the computing complexity and required

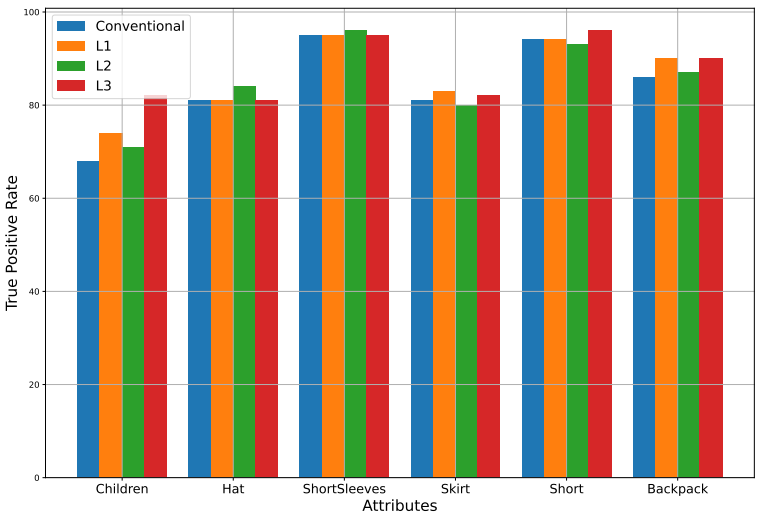

Fig. 4. True positive rate of attributes with unbalanced data.

Tab. 4. Evaluation results of the system on different weight functions.

Method	mA	Acc	Prec	Recall	$F1$
Conventional	94.41 ± 0.04	$\mathbf{91.64 \pm 0.05}$	$\mathbf{94.55 \pm 0.04}$	93.62 ± 0.11	$\mathbf{93.66 \pm 0.04}$
L1 [14]	94.28 ± 0.01	91.57 ± 0.07	94.43 ± 0.07	$\mathbf{93.65 \pm 0.07}$	93.61 ± 0.06
L2 [29]	94.04 ± 0.11	91.47 ± 0.03	94.12 ± 0.04	93.54 ± 0.06	93.39 ± 0.06
L3 ($\alpha = 1$) [36]	$\mathbf{94.85 \pm 0.04}$	91.22 ± 0.09	94.51 ± 0.03	93.36 ± 0.03	93.53 ± 0.02

inference time. In addition, the ConvNeXt architecture uses fewer action functions and normalization layers than ResNet, reducing the computational requirement.

4. Conclusions and outlook

An efficient human attributes recognition system has been successfully presented in this work. Efforts to use a state-of-the-art backbone like Swin-S or ConvNeXt can improve the system's accuracy. However, this improvement is still slight in the range of 1-2%. There is quite a similar trend when optimization techniques like AdamW or RE are used.

Tab. 5. Evaluation results of computational requirements for different models.

Backbone	FLOPs [G]	No. of params [M]	Inference time [ms]
ResNet50 [9]	4.12	23.53	210
Swin-S [21]	8.52	48.85	70
ConvNeXt-v1-B [22]	15.36	87.58	70
ConvNeXt-v2-B [33]	15.36	87.71	80

The use of advanced weight functions can also solve the issue of data unbalancing and thus significantly improve the accuracy. The ConvNeXt-v2-B should be used since it has shown the best accuracy and computational efficiency. Any approaches trying to improve the attribute recognition model are hard to produce remarkable results, especially for practical applications with many challenging conditions. The post-processing technique using keypoints of the pose for filtering attributes proposed in this work is efficient, which improves the system's accuracy significantly. Also, the use of keypoints can make the system robust against the challenging conditions of real-life applications such as images containing occluded or partially visible human body parts. For the best practice system, we should combine two directions. On the one hand, we improve the attribute recognition model with a state-of-the-art backbone, a diverse and large dataset, and optimization training techniques, as shown in this paper. On the other hand, the post-processing technique presented in this work should be added.

Although the keypoints concept has proven to have significant results for the system, it should be further improved. First, the accuracy of the pose detection model can be improved by fine-tuning the Yolov7-Pose on a more diverse and challenging dataset. Second, the algorithm that uses keypoints to filter the attributes must be thoroughly evaluated, especially in real-life and in challenging scenarios. For practical applications, a lightweight model based on the processing pipeline proposed in this work is planned to be developed so that it can be deployed on limited hardware resources of edge computing devices.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Proc. 2011 Int. Conf. Computer Vision (ICCV)*, pages 1543–1550, Barcelona, Spain, 6-13 Nov 2011. IEEE. doi:10.1109/ICCV.2011.6126413.

[2] W.-C. Chen, X.-Y. Yu, and L.-L. Ou. Pedestrian attribute recognition in video surveillance scenarios based on view-attribute attention localization. *Machine Intelligence Research*, 19(2):153–168, 2022. doi:10.1007/s11633-022-1321-8.

[3] X. Cheng, M. Jia, Q. Wang, and J. Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6994–7004, 2022. doi:10.1109/TCSVT.2022.3178144.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA, 20-25 Jun 2009. doi:10.1109/CVPR.2009.5206848.

[5] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proc. 22nd ACM Int. Conf. Multimedia (MM'14)*, ACM Conferences, pages 789–792, Orlando, FL, USA, 3-7 Nov 2014. doi:10.1145/2647868.2654966.

[6] A. Diba, A. M. Pazandeh, H. Pirsiavash, and L. Van Gool. Deepcamp: Deep convolutional action & attribute mid-level patterns. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3565, Las Vegas, NV, USA, 27-30 Jun 2016. doi:10.1109/CVPR.2016.387.

[7] H. Galiyawala, M. S. Raval, and M. Patel. Person retrieval in surveillance videos using attribute recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13, 2022. doi:10.1007/s12652-022-03891-0.

[8] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 2470–2478, Santiago, Chile, 13-16 Dec 2015. doi:10.1109/ICCV.2015.284.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, 27-30 Jun 2016. doi:10.1109/CVPR.2016.90.

[10] J. Jia, H. Huang, X. Chen, and K. Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv*, 2021. arXiv:2107.03576. doi:10.48550/arXiv.2107.03576.

[11] J. Joo, S. Wang, and S.-C. Zhu. Human attribute recognition by rich appearance dictionary. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 721–728, Sydney, Australia, 1-8 Dec 2013. doi:10.1109/ICCV.2013.95.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.

[13] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proc. Workshop on Challenges in Representation Learning (WREPL), part of Int. Conf. Machine Learning (ICML)*, page 896. Atlanta, GE, USA, 16-21 Jun 2013.

[14] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proc. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, Kuala Lumpur, Malaysia, 3-6 Nov 2015. IEEE. doi:10.1109/ACPR.2015.7486476.

[15] D. Li, X. Chen, Z. Zhang, and K. Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *Proc. 2018 IEEE Int. Conf. Multimedia and Expo (ICME)*, pages 1–6, San Diego, CA, USA, 23-27 Jul 2018. doi:10.1109/ICME.2018.8486604.

[16] D. Li, Z. Zhang, X. Chen, and K. Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 28(4):1575–1590, 2018. doi:10.1109/TIP.2018.2878349.

[17] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *Computer Vision, Proc. 14th European Conf. Computer Vision (ECCV 2016)*, volume 9910 Part VI of *Lecture Notes in Computer Science*, pages 684–700, Amsterdam, The Netherlands, 11-14 Oct 2016. Springer. doi:10.1007/978-3-319-46466-4_41.

[18] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. doi:10.1016/j.patcog.2019.06.006.

[19] P. Liu, X. Liu, J. Yan, and J. Shao. Localization guided learning for pedestrian attribute recognition. In *Proc. British Machine Vision Conference (BMVC 2018)*, Northumbria, UK, 3-6 Sep 2018. BMVA Press. Accessible also as arXiv:1808.09102. `https://bmva-archive.org.uk/bmvc/2018/contents/papers/0573.pdf`.

[20] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 350–359, Venice, Italy, 22-29 Oct 2017. doi:10.1109/ICCV.2017.46.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pages 10012–10022, Montreal, QC, Canada, 10-17 Oct 2021. doi:10.1109/ICCV48922.2021.00986.

[22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, New Orleans, LA, USA, 18-24 Jun 2022. doi:10.1109/CVPR52688.2022.01167.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. 7th Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, 6-9 May 2019. `https://openreview.net/forum?id=Bkg6RiCqY7`.

[24] D. Maji, S. Nagori, M. Mathew, and D. Poddar. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2636–2645, New Orleans, LA, USA, 19-20 Jun 2022. doi:10.1109/CVPRW56347.2022.00297.

[25] OpenCV Team. OpenCV, 2022. `https://opencv.org`. [Accessed 15 Jan 2022].

[26] H. X. Nguyen, D. N. Hoang, T. V. Nguyen, T. M. Dang, A. D. Pham, and D.-T. Nguyen. Person re-identification from multiple surveillance cameras combining face and body feature matching. *Modern Physics Letters B*, 37(19):2340031, 2023. doi:10.1142/S0217984923400316.

[27] S. Sakib, K. Deb, P. K. Dhar, and O.-J. Kwon. A framework for pedestrian attribute recognition using deep learning. *Applied Sciences*, 12(2):622, 2022. doi:10.3390/app12020622.

[28] A. Specker, M. Cormier, and J. Beyerer. UPAR: Unified Pedestrian Attribute Recognition and person retrieval. In *Proc. 2023 IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, pages 981–990, Los Alamitos, CA, USA, 3-7 Jan 2023. doi:10.1109/WACV56688.2023.00104.

[29] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proc. AAAI Conf. Artificial Intelligence*, volume 34 of *AAAI-20 Technical Tracks 7*, pages 12055–12062, New York, NY, USA, 7-12 Feb 2020. AAAI Press. doi:10.1609/aaai.v34i07.6883.

[30] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, Vancouver, Canada, 18-22 Jun 2023. Accessible also as arXiv:2207.02696. `https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html`.

[31] X. Wang, S. Zheng, R. Yang, A. Zheng, Z. Chen, J. Tang, and B. Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022. doi:10.1016/j.patcog.2021.108220.

[32] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, Salt Lake City, UT, USA, 18-23 Jun 2018. doi:10.1109/CVPR.2018.00016.

[33] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, 18-22 Jun 2023. Accessible also as arXiv:2301.00808. `https://openaccess.thecvf.com/content/CVPR2023/html/Woo_ConvNeXt_V2_Co-Designing_and_Scaling_ConvNets_With_Masked_Autoencoders_CVPR_2023_paper.html`.

[34] L. Yang, L. Zhu, Y. Wei, S. Liang, and P. Tan. Attribute recognition from adaptive parts. *arXiv*, 2016. arXiv:1607.01437. doi:10.48550/arXiv.1607.01437.

[35] N. Zhang, M. Paluri, M'A. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose Aligned Networks for Deep Attribute modeling. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1644, Columbus, OH, USA, 23-28 Jun 2014. doi:10.1109/CVPR.2014.212.

[36] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2370, Nashville, TN, USA, 20-25 Jun 2021. doi:10.1109/CVPR46437.2021.00239.

[37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proc. AAAI Conf. Artificial Intelligence*, volume 34 of *AAAI-20 Technical Tracks 7*, pages 13001–13008, New York, NY, USA, 7-12 Feb 2020. AAAI Press. doi:10.1609/aaai.v34i07.7000.

[38] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *Proc. 2015 Int. Conf. Biometrics (ICB)*, pages 535–540, Phuket, Thailand, 19-22 May 2015. IEEE. doi:10.1109/ICB.2015.7139070.

Ha X. Nguyen received his Ph.D. degree in computing science and micro-robotics from the University of Oldenburg, Germany, in 2014. He is now working as a lecturer for intelligent robotics at Hanoi University of Science and Technology. He also serves as a consultant expert at the IoT/Smart-Devices Laboratory at the CMC Applied Technology Institute, CMC Corporation. His research interests cover intelligent robots, micro-robotics, and computer vision.

Dong N. Hoang has five years of research experience in machine learning, computer vision, and building security surveillance systems. His research fields include facial, human, vehicle, and building scalable pipeline architectures. He received his B.Sc. degree in 2015 from the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam. Now he is working as the deputy head of the IoT/Smart Device Laboratory at the CMC Applied Technology Institute, CMC Corporation, Hanoi, Vietnam.

Tuan A. Tran has three years of research experience in machine learning, computer vision, and optimization. He received his B.Sc. degree in 2022 from School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, Vietnam. Now, he is working as researcher at CMC Applied Technology Institute, CMC Corporation, Hanoi, Vietnam.

Tuan M. Dang received his Ph.D. degree in mathematics at Academy of Military Science and Technology, Ministry of Defense, Vietnam. He is now working as a lecturer for computer science at Posts and Telecommunication Institute of Technology, Vietnam, and researcher at CMC Applied Technology Institute, CMC Corporation, Hanoi, Vietnam. His research interests cover cryptography, blockchain and artificial intelligence.