CrossPAR: Enhancing Pedestrian Attribute Recognition with Vision-Language Fusion and Human-Centric Pre-training

Bach-Hoang Ngo[1,2][0009−0002−2290−1187], Si-Tri Ngo[1,2], Phu-Duc Le[1,2], Quang-Minh Phan[1,2], Minh-Triet Tran[1,2][0000−0003−3046−3041], and Trung-Nghia Le*[1,2][0000−0002−7363−2610]

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam {nhbach22, ntsi22, ldphu22, pmquang22}@apcs.fitus.edu.vn, {tmtriet, ltnghia}@fit.hcmus.edu.vn

Abstract. Pedestrian attribute recognition (PAR) is crucial in various applications like surveillance and urban planning. Accurately identifying attributes in diverse and intricate urban environments is challenging despite its significance. This paper introduces a novel network for PAR that integrates a human-centric encoder, trained on extensive human datasets, with a vision-language encoder, trained on substantial text-image pair datasets. We also develop a cross-attention mechanism utilizing a Mixture-of-Experts approach that combines the human-centric encoder's proficiency in local attribute detection with the vision-language encoder's ability to comprehend global content. CrossPAR showcases a comparable accuracy result compared to existing techniques across multiple benchmarks, using less training data. These results confirm our approach's effectiveness and suggest promising avenues for further research and practical applications in the domain of PAR and related fields.

1 Introduction

In the evolving landscape of artificial intelligence, *pedestrian attribute recognition* (PAR) has emerged as an essential task with wide-ranging applications, such as pedestrian re-identification and retrieval [32], autonomous vehicle technology [22], and video surveillance [15]. At its core, PAR involves identifying and classifying various attributes of pedestrians, including age, gender, and clothing [12]. Numerous significant advancements in PAR have been made in recent years [3, 4, 11, 28, 29]. However, PAR is fraught with significant challenges. Low-quality images obtained from surveillance cameras [19] can significantly hinder the accuracy of attribute detection. Occlusions, where objects or other individuals partially obscure pedestrians, add another layer of complexity, making it challenging to identify attributes accurately. Furthermore, the diverse range of

* Corresponding author.

environments where pedestrians are captured [5, 15, 19] introduces variability that can confuse even the most sophisticated algorithms.

Conventional PAR methods have mainly relied on utilizing the capabilities of single-image encoders, trained either on large-scale image datasets such as ImageNet [11, 26] or those trained on datasets specifically tailored for human-centric tasks [3, 4, 29]. However, this strategy may not be the most effective. The development of text-to-image trained encoders [10, 23, 25, 30, 33], particularly those refined through pedestrian retrieval datasets [1, 13, 32], introduces a promising avenue. These models excel at analyzing the intricate relationships between various attributes, providing a solid foundation for classification tasks. Their enhanced understanding could improve performance in challenging scenarios, such as when dealing with occlusions or images of low quality.

In addition, the issue of effectively representing both local and global features within an image has emerged as a concerning area of focus in PAR [7]. The balance between capturing the minor details (local features) that define individual attributes and understanding the broader context (global features) within which these attributes exist is essential. This challenge leads to the interest in hybrid models [21], which combine the strengths of diverse networks to achieve a more holistic understanding of the image. Such models have already begun to show promising results, pointing towards a new frontier in PAR research.

Building on these premises, we explore the integration of text-to-image models. By leveraging their capability of capturing attribute relationships, there's potential to elevate the accuracy of attribute classification and extend the robustness of these systems under less-than-ideal conditions, such as occlusion or low-quality pictures. To this end, we introduce a novel PAR framework called CrossPAR, integrating advanced transformer-based models, specifically a Swin Transformer [20] architecture pre-trained on large-scale pedestrian dataset [3] and a Vision Transformer [33] based architecture fine-tuned on text-based pedestrian retrieval data [17]. We aim to leverage these models' inherent strengths in processing complex visual data. We also propose a cross-attention module with a mixture of experts to merge the representations derived from these models. Our cross-mixture-of-experts module can fully take advantage of the unique capabilities of the different encoders, leading to a more robust and accurate PAR system, outperforming existing methods by 1-2% in the PA100K dataset [19]. Our contributions are delineated as follows:

- We propose CrossPAR, a network that capitalizes on the unique advantages offered by diverse transformer architectures, specifically tailored toward different human-centric objectives.
- We investigate vision-language and human-centric encoders to harness their sophisticated capability for complex representation.
- We introduce a cross-mixture-of-experts block, which can facilitate enhanced interaction between the different elements for merging components in our hybrid network. We also undertake a comprehensive exploration of different fusion algorithms.

The remainder of this paper is structured as follows: Sec. 2 discusses the related works leading to this research. Sec. 3 describes the proposed method by investigating the benefits of using text-image and human-centric encoder. Sec. 4 shows and discusses the performance of the proposed system. Finally, Sec. 5 presents the concluding remarks and mentions our future work.

2 Related Work

2.1 Pedestrian Attribute Recognition (PAR)

In the realm of PAR, a primary hurdle lies in the imbalanced distribution of human attributes, which undermines the efficacy of PAR models. Li et al. [16] addressed this challenge by advocating the adoption of a weighted binary cross-entropy loss function. Building upon this, Liu et al. [19] introduced a multi-directional attention module that capitalizes on multi-scale feature maps to enrich the final feature representation. Tang et al. [28] proposed a comprehensive framework that combines Feature Pyramid Network (FPN), Spatial Transformer Network (STN), and Squeeze-and-Excitation (SE) modules to refine localization accuracy. Subsequently, Jia et al. [12] reevaluated PAR, redefining the approach and presenting an explicit framework facilitating accurate attribute prediction for unseen pedestrians. Inspired by their approach, we introduce a novel cross-attention fusion that facilitates earlier feature interaction.

2.2 Pedestrian Vision-Language Models

The integration of vision-language pre-training has garnered substantial interest in recent years [23,25], opening up promising avenues across a spectrum of computer vision tasks. Notably, the CLIP model [23] has emerged as a pioneering framework in this domain, laying the groundwork for subsequent advancements in vision-language fusion. Subsequently, various methods have been developed to enhance the alignment between textual and visual representations [18,25,33]. This concerted effort aims to refine the synergy between language and vision modalities, thereby improving the overall interpretability and effectiveness of models. Using language as a foundational element in vision tasks has thus become increasingly compelling, offering new avenues for enhanced comprehension and interpretation of visual content [8].

Building upon this trend, several pedestrian vision-language datasets [6, 17] and methods [1, 14, 32] have been proposed. Initially developed for text-based pedestrian retrieval, these models can leverage the synergy between vision and language modalities to understand and interpret pedestrian attributes. However, recognizing their potential beyond text-based retrieval, we acknowledge these models as robust candidates for fine-tuning in PAR. Their robust pre-trained knowledge of pedestrian attributes, acquired through comprehensive text captions, positions them as valuable assets for enhancing the accuracy and efficiency of attribute recognition systems.

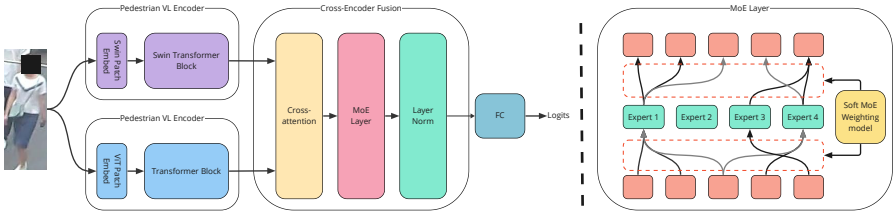

Fig. 1: Overview architecture of the proposed CrossPAR. Our method employs a dual-branch pipeline integrating human-centric and pedestrian vision-language encoders. It utilizes Swin Transformer Block for human-focused features in one branch and a Transformer Block for pedestrian-specific visual language features in the other, refined through multi-head self-attention and merged via a Cross-Encoder Fusion module.

2.3 Feature Fusion

Utilizing a cross-encoder architecture with cross-attention blocks is a notable strategy in the text-to-image retrieval sector [14, 33]. These blocks act as a pivotal fusion layer, adept at integrating embeddings from varied modalities, thereby enhancing the interpretative depth of the models. Despite its prevalent application in cross-modal retrieval tasks, the exploration of this approach for in-modality feature fusion within PAR domain remains limited. An exception to this is the introduction of a cross-fuse block by Bui et al. [2]. However, their methodology did not extend to employing the cross-attention module for the joint training of image encoders, leaving room for further exploration.

Motivated by this gap, we adopt the cross-attention block, leveraging it as an advanced ensemble technique to bolster the synergy among various image encoders. This methodology aims to augment their collective performance in PAR, taking advantage of the cross-attention mechanism's capacity to prioritize relevant image features dynamically. This focus adjustment, guided by text and additional visual signals, enables a deeper and more detailed analysis of images. It boosts the system's ability to identify and understand complex attributes with enhanced accuracy.

3 Proposed Method

3.1 Overview

An overview of our pipeline can be seen in Fig. 1. Using a dual-branch pipeline, we propose the architecture integrating a human-centric approach and a pedestrian vision-language (VL) encoder. In the human-centric branch, we utilize a Swin Transformer-based backbone [3], trained on general human-centric tasks, to extract features focusing on humans. Conversely, in the pedestrian VL branch, we employ a Transformer-based encoder [33] to extract pedestrian-specific visual language features. These features are refined through multi-head self-attention in

both branches before being combined via a Cross Fusion module. Subsequently, the merged features undergo projection and processing through a fully connected (FC) layer to generate final logits, facilitating PAR.

3.2 Human-centric Encoder

In this paper, we propose leveraging a human-centric encoder for PAR, pre-trained on a diverse set of human-centric data. The rationale behind this approach lies in the inherent capability of human-centric models to capture fine-grained details and semantic features for a crucial understanding of human attributes in visual data.

We employ the state-of-the-art human-centric encoder SOLIDER [3], designed to learn comprehensive human representations from unlabeled human images. Unlike traditional self-supervised approaches that primarily focus on appearance features, SOLIDER integrates semantic information into learning by leveraging pseudo-semantic labels derived from prior knowledge of human images. This inclusion of semantic details allows the framework to generate more contextually rich representations suitable for a wide array of human-centric tasks, including PAR.

By utilizing a human-centric encoder for PAR, we aim to capitalize on its innate ability to comprehend human-centric visual cues, thereby potentially improving the accuracy and efficiency of PAR systems in real-world scenarios.

3.3 Pedestrian Vision-Language Backbone

We incorporate Vision Language Models (VLMs) [33] trained on large-scale text-to-image pedestrian retrieval dataset CUHK-PEDES [17] into our pipeline for PAR. By encoding textual descriptions associated with pedestrian images, VLMs offer a deeper semantic understanding of pedestrian attributes, capturing fine-grained details and contextual cues crucial for accurate recognition. Through cross-modal learning, VLMs align textual descriptions with visual features, enabling the model to correlate linguistic attributes with corresponding visual representations, thus improving the granularity and specificity of attribute recognition. Additionally, these encoders enhance the robustness of attribute recognition systems by mitigating ambiguities in visual data and generalizing well to unseen attributes. Overall, incorporating VLMs enriches the PAR process, leading to more precise, contextually aware attribute recognition systems with improved performance and generalization capabilities.

3.4 Cross-Encoder Fusion

Drawing inspiration from prior research in multimodal learning [14,31,33], we've developed a new component called the Cross-Encoder Fusion Block to merge the strengths of two main frameworks effectively: the Human-centric encoder and Vision Language Models. This innovative block utilizes cross-attention mechanisms, allowing it to blend the visual and textual information extracted by the

encoders. By doing so, it can dynamically highlight important parts and aspects of an image while integrating text-based context.

To further improve the fusion capability of the cross-attention block, we proposed the integration of a mixture-of-experts (MoE) block to fuse the image patches. This approach allows the model to leverage multiple expert networks, each tailored to different aspects of the input data, thereby enhancing the overall processing of pedestrian features. The MoE block dynamically selects the most relevant experts for each pedestrian segment. This selection is based on the content of the image patches and the unique competencies of each expert, ensuring that the most effective combinations of features are used for decision-making. By distributing the computational load across various experts, the MoE block improves the model's accuracy and efficiency in handling complex image data. This modular approach is particularly advantageous in challenging situations, such as occlusions or low-quality images, where conventional processing might fail to deliver precise results.

3.5 Loss Function

Following the precedent set by prior study [12], we utilize a binary cross-entropy loss function for multi-class classification. Here, $\hat{y} \in R^{N_{attr}}$ represents the output from the classification head of the fusion block, post application of the *Sigmoid* function, while y denotes the ground truth in one-hot encoding format. The binary cross-entropy loss function $\mathcal{L}_{BCE}$ is defined as:

$$\mathcal{L}_{BCE}(\hat{y}, y) = -(ylog(\hat{y}) + (1-y)log(1-\hat{y})). \tag{1}$$

In line with enhancing pipeline performance, we extend the binary cross-entropy (BCE) loss computation to encompass both the Human-centric encoder and the Pedestrian Vision-Language Encoder outputs. The final loss is a cumulative sum of three components:

$$\mathcal{L}_{total} = \mathcal{L}_H + \mathcal{L}_{VL} + \mathcal{L}_{fuse}, \tag{2}$$

where $\mathcal{L}_H$ represents the BCE loss for the human-centric encoder, $\mathcal{L}_{VL}$ denotes the BCE loss for the vision-language encoder and $\mathcal{L}_{fuse}$ stands for the loss associated with the fusion block. This comprehensive approach ensures robust training across all encoder components, contributing to the system's overall effectiveness.

4 Experiments

4.1 Implementation Details

In our implementation, we synergistically combine a human-centric encoder with a vision-language encoder to enhance our experimental framework. In this work, we employed the SOLIDER architecture [3] as the foundational human-centric encoder based on the Swin Transformer Base (SwinT-B) with an input resolution of 224×224. This particular image encoder, pre-trained on an extensive

human dataset encompassing over 4.18 million data samples, was selected for its proven capability to capture human-centric features with high fidelity. Concurrently, the integrate vision and language modalities, the X2-vlm [33] was utilized. This model, fine-tuned on the large text-image CUHK-PEDES dataset [17], was instrumental in facilitating a nuanced interpretation of complex visual-linguistic interactions. In our experiments, we set the learning rate to 1e-5, employing a cosine learning rate scheduler in conjunction with the AdamW optimizer. The input resolution for images was standardized at 224×224.

4.2 Experimental Settings

Dataset All experiments were carried out using the PA-100K dataset [19], a comprehensive collection of 100,000 pedestrian images spanning 598 distinct scenes. The dataset encompasses 26 attributes, capturing various pedestrian characteristics such as handbag possession, phone usage, gender, age, etc. This dataset is sourced from many camera setups, ensuring a broad representation of pedestrian appearances and scenarios.

Evaluation Metrics To assess the performance of attribute recognition, following prior work [3,4,24,27,29], we employ two categories of metrics: a label-based metric and four instance-based metrics. The label-based metric involves calculating the average classification accuracy for both positive and negative samples for each attribute. This average, known as **mean Accuracy (mA)**, is then computed across all attributes. On the other hand, the instance-based metrics consist of **Accuracy (InsAcc)** and **F1-score (InsF1)**, providing a comprehensive evaluation of the system's performance.

4.3 Comparison with State-of-the-Art Methods

We evaluate our method on the PA-100K benchmark against a range of state-of-the-art approaches [3,4,9,24,27–29]. The comparison evaluates each method based on its encoder architecture and performance metrics, specifically mA and F1-score, which are crucial for assessing the effectiveness of PAR models.

The comparative analysis in Tab. 1 shows that our CrossPAR, leveraging the fusion of the human-centric encoder and vision-language encoder, is a robust and competitive approach for PAR on the PA-100K dataset. CrossPAR demonstrates competitive performance with an mA of 86.9 and an InsF1 of 90.6. Notably, our method achieved the highest values of mA and InsF1, outperforming other methods. The high value of InsF1 also indicates the effectiveness of CrossPAR in maintaining a balanced precision-recall trade-off, which is essential for practical applications where both false positives and false negatives carry significant consequences.

The state-of-the-art PATH [29], utilized a combination of 6 PAR datasets in their training process, amounting to a total of 242,880 images specifically for the PAR task, alongside additional data for other tasks. In comparison, our

Table 1: Comparison of our CrossPAR against state-of-the-art methods. ↑ means higher is better.

Method	Publication	Backbone	mA ↑	InsF1 ↑
VAC [9]	CVPR 2019	ResNet-50	79.0	86.8
ALM [28]	ICCV 2019	BN-Inception	80.7	86.5
JLAC [27]	AAAI 2020	ResNet-50	82.3	87.6
UPAR [24]	WACV 2022	ConvNeXt-B	84.8	90.2
UniHCP [4]	CVPR 2023	Enc-Dec	86.2	-
SOLIDER [3]	CVPR 2023	SwinT-B	86.4	-
PATH [29]	CVPR 2023	ViT-B	86.9	-
CrossPAR (ours)	ACCV 2024	**SwinT-S + x2vlm**	**86.9**	**90.6**

Table 2: Performance of different encoder architectures on PA 100K Dataset, blue denoted the best-performing encoder without fusion.

Encoder	mA	InsAcc	InsF1
ResNet50	67.9	61.1	74.6
Swin-B 224	83.9	80.5	87.8
ViT-B	81.6	79.2	87.5
ConvNeXt Base	83.9	81.1	88.8
EfficientNet B3	78.2	74.3	84.3
Human-centric (HC)	85.4	84.1	90.6
Vision language (VL)	85.6	82.8	89.7
HC + VL	**86.4**	**84.2**	**90.6**

method was trained using only the PA-100K dataset, which has 80,000 training images for the PAR task, yet we achieved the same results (86.9 in mA) with less training data.

4.4 Ablation Study

Impact of Human-Centric and Vision-Language Encoders. As seen in Tab. 2, both human-centric and vision language encoders perform better than other architectures on the PA 100K dataset. The vision language encoder achieves the highest mean accuracy (mA) among the non-fusion methods at 85.6%, closely followed by the human-centric encoder at 85.5%. However, the human-centric encoder has a slight advantage regarding instance-level F1-score (InsF1) at 90.6%

The experimental result suggests that while both human-centric and vision language encoders show excellent performance in recognizing pedestrian attributes, they focus on slightly different aspects of the task. Human-centric encoders might better capture detailed information important for accurate attribute classification (higher InsF1), while vision language encoders are great at using semantic relationships for overall attribute prediction (higher mA). Com-

Table 3: Results of fusion algorithms.

Fusion Method	mA	InsAcc	InsPrec	InsRecall	InsF1	Mean
Concat	85.0	81.5	88.4	89.5	88.4	86.7
Add	85.5	82.2	88.6	89.8	89.1	87.1
Cross Attention	86.4	**84.2**	90.51	90.73	90.6	88.5
Cross MoE	**86.9**	82.9	**91.8**	**90.9**	90.2	**88.6**

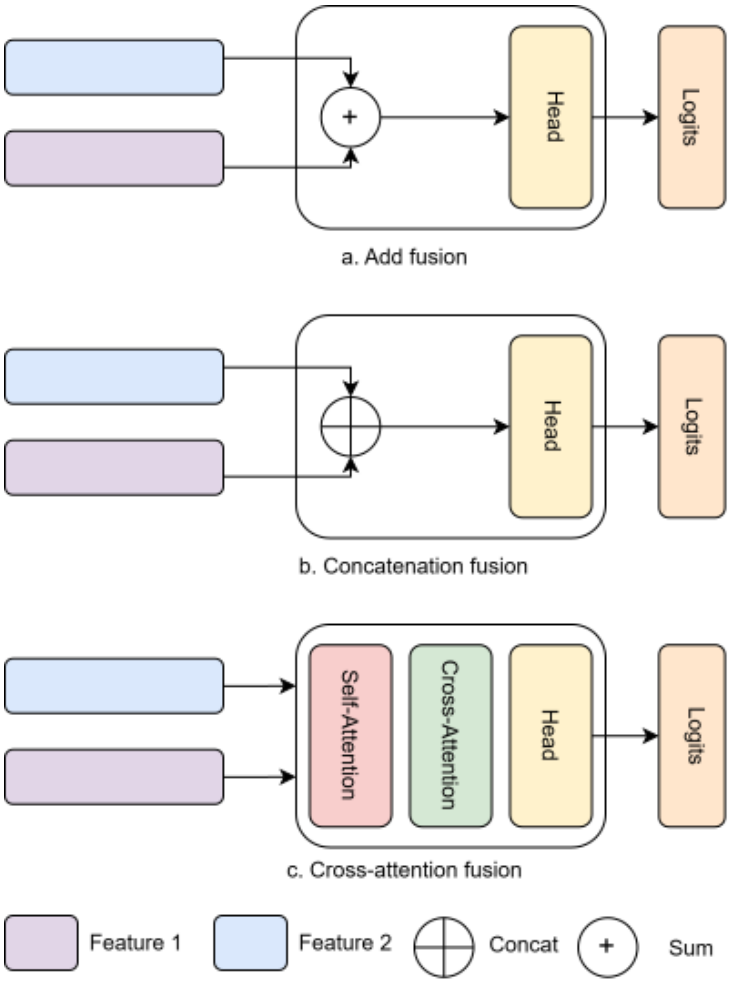

Fig. 2: Evaluated fusion algorithms.

bining these two encoders enhances overall performance, leveraging the detailed attribute classification strengths of human-centric encoders and the semantic relationship proficiency of vision language encoders.

Effectiveness of Fusion Algorithms. We explored three primary fusion algorithms to combine two encoders: addition, concatenation, and cross-attention (see Fig. 2). Through experimental results in Tab. 3, we observed notable variations in performance across these fusion algorithms.

The performance of the concatenation and addition fusion algorithms was closely examined for their efficacy in synthesizing and interpreting complex multimodal information. Despite its utility, the concatenation method registered the lowest performance, indicating room for improvement in feature integration. On the other hand, the addition method showed a subtle improvement, suggesting a better, yet not optimal, fusion of features. These results highlight the inherent limitations of both simple fusion methods in achieving the most effective and precise attribute recognition in complex datasets.

However, a breakthrough was observed with the implementation of the cross-attention fusion method. This advanced approach significantly outperformed the methods above by achieving the highest values of InsAcc and InsF1, reflecting its ability to identify relevant attributes but also its efficiency in minimizing false negatives, thereby ensuring a comprehensive and accurate attribute recognition process.

On the other hand, the cross-mixture-of-experts method achieved the highest mA, underscoring the unparalleled efficiency of our fusion method. This approach's standout performance is attributed to its sophisticated mechanism of focusing on and integrating relevant features from the multimodal input, thereby facilitating a more nuanced and accurate interpretation of the data. This approach fosters a powerful collaboration between the visual and textual elements, leading to improved recognition of attributes and a deeper understanding of semantics. Its adoption can significantly enhance the interpretive capabilities of models, paving the way for more accurate and reliable analyses in diverse applications.

Impact of Number of Experts. We analyze the impact of varying the number of experts in the MoE layer on the model's performance. The performance metric is plotted against ensemble sizes of 4, 8, 12, and 16 experts in Fig. 3. There is an upward trend in performance as the number of experts increases, with the model achieving the highest score of 86.9% with 16 experts. However, the performance differences between the configurations are marginal, with scores ranging from 86.5% to 86.9%. This suggests that the gains are not substantial, while scaling up the number of experts can lead to slight improvements. Therefore, staying at 4 experts could be a good choice for maintaining a good performance while keeping the model computational cost relatively controllable.

4.5 Failure Cases

In the PA-100K dataset, occlusions and low image quality present formidable challenges for pedestrian recognition models. Occlusions, as shown in some examples in Fig. 4a, such as text overlays obscure vital pedestrian features. Low-quality images with indistinct or barely visible pedestrians, as shown in Fig. 4b,

Fig. 3: Impact of the number of experts in the MoE layer.

hinder accurate identification. These obstacles demand robust algorithms that handle noisy, incomplete, or degraded visual inputs. Solutions may involve advanced image processing techniques and dataset augmentation strategies to enhance model adaptability and performance. Overcoming these challenges is crucial for advancing pedestrian recognition technology and its applications in real-world scenarios.

In analyzing clear images without occlusions, we compare the Human-Centric (HC) and Vision-Language (VL) models, focusing on samples with the lowest scores. A notable trend of confusion arises, primarily between front, back, and side attributes. The HC model excels in recognizing visible attributes like gender, age, and clothing types, while the VL model performs better with context-dependent attributes like Backpack, Side, and Front. Many instances exhibit misclassifications where "front" is mistaken for "back" or "side," largely due to inherent ambiguity. Additionally, sleeve length is frequently misclassified, with confusion between "long sleeve" and "short sleeve." This ambiguity is apparent in Fig. 5, where accuracy for these attributes is lower than average. Enhancing model robustness to subtle visual cues and combining both models could refine pedestrian attribute recognition by leveraging each model's strengths.

5 Conclusion

This paper presents CrossPAR, a novel network that strategically leverages the strengths of various transformer architectures; each is chosen for its efficacy in addressing distinct aspects of human-centric analysis. By integrating a sophisticated text-image, human-centric encoder, our approach achieves a nuanced and complex representation essential for understanding the intricacies of human attributes in images. In addition, our introduced cross-attention with MoE proves

(a) Occlusion cases in the PA-100K dataset

(b) Low image quality examples in the PA-100K dataset

Fig. 4: Examples of occlusion and low image quality in the PA-100K dataset

instrumental in fostering dynamic interactions between disparate network elements, thereby enriching the model's interpretative capabilities. Rigorous evaluations conducted on many PAR benchmarks underscore the robustness and versatility of our proposed method, affirming its potential to set new precedents in the realm of PAR.

6 Potential Negative Society Impact

PAR offers valuable insights for various applications, such as identifying suspects or track individuals in public spaces, monitoring pedestrian flow, optimizing traffic management, analyzing customer demographics to tailor marketing strategies. But it also poses significant challenges to individual privacy. Attributes can be linked back to an individual, thereby raising concerns about the invasion of privacy. The data collected by PAR systems are vulnerable to security breaches or can be misused for malicious purposes. To mitigate such risks, organizations deploying PAR systems should be transparent about their use and collected

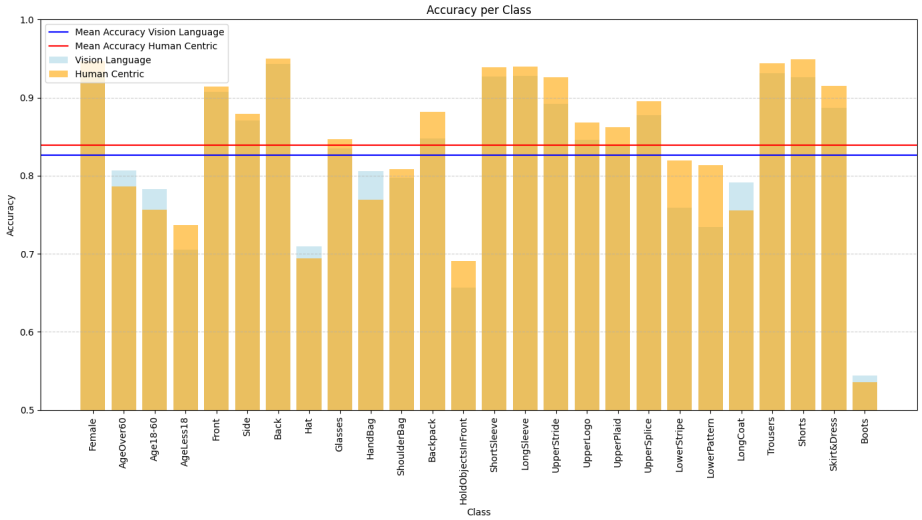

Fig. 5: Result of model's accuracy for each class, with notable success observed in categories associated with clothing and age groups, while those linked to items and directions tend to fall below the mean accuracy threshold.

data. Developers of PAR should prioritize privacy-preserving techniques, such as anonymization and differential privacy, to minimize the risks associated with data collection.

Acknowledgement. This research is funded by University of Science, VNU-HCM, under grant number CNTT 2024-16.

References

1. Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: Rasa: Relation and sensitivity aware representation learning for text-based person search. arXiv preprint arXiv:2305.13653 (2023) 2, 3
2. Bui, D.C., Le, T.V., Ngo, B.H.: C2t-net: Channel-aware cross-fused transformer-style networks for pedestrian attribute recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 351–358 (January 2024) 4
3. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15050–15061 (June 2023) 1, 2, 4, 5, 6, 7, 8
4. Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17840–17852 (June 2023) 1, 2, 7, 8

5. DENG, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia. p. 789–792. MM '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2647868.2654966, https://doi.org/10.1145/2647868.2654966 2
6. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021) 3
7. Djenouri, Y., Belbachir, A.N.: A hybrid visual transformer for efficient deep human activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 721–730 (October 2023) 2
8. Dong, X., Bao, J., Zhang, T., Chen, D., Shuyang, G., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Clip itself is a strong fine-tuner: Achieving 85.788.0arXiv:2212.06138 (2022) 3
9. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 7, 8
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. ArXiv **abs/2102.05918** (2021), https://api.semanticscholar.org/CorpusID:231879586 2
11. Jia, J., Chen, X., Huang, K.: Spatial and semantic consistency regularizations for pedestrian attribute recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 962–971 (2021) 1, 2
12. Jia, J., Huang, H., Chen, X., Huang, K.: Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. arXiv preprint arXiv:2107.03576 (2021) 1, 3, 6
13. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2
14. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3, 4, 5
15. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition. ArXiv **abs/1603.07054** (2016) 1, 2
16. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) pp. 111–115 (2015), https://api.semanticscholar.org/CorpusID:9475404 3
17. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. arXiv preprint arXiv:1702.05729 (2017) 2, 3, 5, 7
18. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=zq1iJkNk3uN 3
19. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 1, 2, 3, 7

20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
21. Liu, Z., Zhang, Z., Li, D., Zhang, P., Shan, C.: Dual-branch self-attention network for pedestrian attribute recognition. Pattern Recogn. Lett. p. 112–120 (nov 2022). https://doi.org/10.1016/j.patrec.2022.10.003 2
22. Mordan, T., Cord, M., Pérez, P., Alahi, A.: Detecting 32 pedestrian attributes for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems (T-ITS) (2021). https://doi.org/10.1109/TITS.2021.3107587 1
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 2, 3
24. Specker, A., Cormier, M., Beyerer, J.: Upar: Unified pedestrian attribute recognition and person retrieval. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 981–990 (2022), https://api.semanticscholar.org/CorpusID:252090333 7, 8
25. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023) 2, 3
26. Tan, Z., Yang, Y., Wan, J., Guo, G., Li, S.Z.: Relation-aware pedestrian attribute recognition with graph convolutional networks **34**, 12055–12062 (Apr 2020) 2
27. Tan, Z., Yang, Y., Wan, J., Guo, G., Li, S.Z.: Relation-aware pedestrian attribute recognition with graph convolutional networks. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 12055–12062 (Apr 2020). https://doi.org/10.1609/aaai.v34i07.6883, https://ojs.aaai.org/index.php/AAAI/article/view/6883 7, 8
28. Tang, C., Sheng, L., Zhang, Z., Hu, X.: Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4997–5006 (2019) 1, 3, 7, 8
29. Tang, S., Chen, C., Xie, Q., Chen, M., Wang, Y., Ci, Y., Bai, L., Zhu, F., Yang, H., Yi, L., Zhao, R., Ouyang, W.: Humanbench: Towards general human-centric perception with projector assisted pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21970–21982 (June 2023) 1, 2, 7, 8
30. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19175–19186 (June 2023) 2
31. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 5
32. Yang, S., Zhou, Y., Wang, Y., Wu, Y., Zhu, L., Zheng, Z.: Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In: Proceedings of the 2023 ACM on Multimedia Conference (2023) 1, 2, 3
33. Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., Zhou, W.: X2-vlm: All-in-one pre-trained model for vision-language tasks. arXiv preprint arXiv:2211.12402 (2022) 2, 3, 4, 5, 7