UPAR Challenge 2024: Pedestrian Attribute Recognition and Attribute-based Person Retrieval - Dataset, Design, and Results

Mickael Cormier[1,2,3] Andreas Specker[1,2,3] Julio C. S. Jacques Junior[7,4]
Lennart Moritz[2] Jürgen Metzler[2,3] Thomas B. Moeslund [5]
Kamal Nasrollahi[5,6] Sergio Escalera[4,5,7] Jürgen Beyerer[1,2,3]

[1]Fraunhofer IOSB, Germany, {firstname.lastname}@iosb.fraunhofer.de
[2]Karlsruhe Institute of Technology, Germany, {firstname.lastname}@kit.edu
[3]Fraunhofer Center for Machine Learning, Germany
[4]Computer Vision Center, Spain
[5]Aalborg University, Denmark, {kn,tbm}@create.aau.dk
[6]Milestone Systems, Denmark, kna@milestone.dk
[7]University of Barcelona, Spain, julio.silveira@ub.edu, sergio@maia.ub.es

Abstract

Attribute-based person retrieval enables individuals to be searched and retrieved using their soft biometric features, for instance, gender, accessories, and clothing colors. The process has numerous practical use cases, such as surveillance, retail, or smart cities. Notably, attribute-based person retrieval empowers law enforcement agencies to efficiently comb through vast volumes of surveillance footage from extensive multi-camera networks, facilitating the swift localization of missing persons or criminals. However, for real-world application, attribute-based person retrieval is required to generalize to multiple settings in indoor and outdoor scenarios with their respective challenges. For its second edition, the WACV 2024 Pedestrian Attribute Recognition and Attribute-based Person Retrieval Challenge (UPAR-Challenge) aimed once again to spotlight the current challenges and limitations of existing methods to bridge the domain gaps in real-world surveillance contexts. Analogous to the first edition, two tracks are offered: pedestrian attribute recognition and attribute-based person retrieval. The UPAR-Challenge 2024 dataset extends the UPAR dataset with the introduction of harmonized annotations for the MEVID dataset, which is used as a novel test domain. To this aim, 1.1M additional annotations were manually labeled and validated. Each track evaluates the robustness of the competing methods to domain shifts by training and evaluating on data from entirely different domains. The challenge attracted 82 registered participants, which was considered a success from the organizers' perspective. While ten competing teams surpassed the baseline for track 1, no team managed to outperform the baseline on track 2, emphasizing the task's difficulty. This work describes the challenge design, the adopted dataset, obtained results, as well as future directions on the topic. The UPAR-Challenge dataset is available on GitHub: https://github.com/speckean/upar_challenge.

Figure 1. **PAR datasets –** Sample images from the four sub-datasets contained in the UPAR challenge 2024 dataset. Each dataset shows different characteristics and thus poses different challenges. Market1501 [49], PA100k [26], and PETA [8] are employed for training and validation, and images from the MEVID [7] dataset serve as test images.

1. Introduction

Searching extensive person image databases for individuals matching specific sets of semantic attributes is facilitated by attribute-based retrieval systems. These systems often rely on Person Attribute Recognition (PAR) methods to recognize semantic attributes of individuals, such as gender, age, or clothing information. However, performing PAR and attribute-based person retrieval in surveillance data poses challenges, particularly within single domains. These challenges arise from limitations in image quality, localized attributes, and visibility issues caused by varying viewing angles or occlusions.

To ensure the successful deployment and long-term use of machine learning algorithms in surveillance contexts, it becomes crucial to address the robustness of these algorithms to domain gaps that occur due to changes in the environment. This adaptability is vital for maintaining the effectiveness of the algorithms over time and in diverse scenarios.

Performing generalization experiments for PAR has been a challenging task, primarily due to the lack of a large dataset specifically designed for this purpose. Until recently, researchers faced difficulties in finding a suitable dataset that could adequately support generalization experiments. The absence of such a dataset limited the ability to assess the performance and effectiveness of their models in real-world scenarios. However, with the introduction of the UPAR dataset [5, 37], there is a large-scale dataset specifically curated for generalization experiments. This dataset and its first challenge extension provide a more comprehensive and diverse sample of data with the harmonization of 40 binary attributes over 12 attribute categories across five existing datasets, allowing researchers to evaluate the robustness and adaptability of algorithms across different domains, environments, and scenarios. The availability of these datasets has opened up new opportunities for studying and improving the generalization capabilities of machine learning algorithms, ultimately leading to more reliable and effective models.

For its second edition, the WACV 2024 Pedestrian Attribute Recognition and Attributed-based Person Retrieval Challenge (UPAR-Challenge)[1] aims to demonstrate that the problem of domain gaps in a real-world surveillance context and the related challenges and limitations of existing methods remain and show the progress made in the last year in this field. The problem of domain shifts is particularly present when only limited training data is available and when the test data follows a different inherent data or attribute distribution. Exemplary images are provided in Fig. 1. In this year's challenge, Market1501 [23, 49], PA100k [26], and PETA [8] are employed for training and validation, and images from the MEVID [7] dataset serve as test images. To this aim, a total of 1.1M new binary annotations are contributed to a subset of images sampled from MEVID.

The UPAR Challenge 2024 is split into two tracks associated with semantic pedestrian attributes, such as gender or clothing information: Person Attribute Recognition (PAR) and attribute-based person retrieval. Both tracks are evaluated with public and private sets, with the aim of testing the robustness of the competing methods to domain shifts by training on data from multiple domains and generalizing to an unknown target domain. The challenge attracted a total of 82 registered participants in its different tracks. With a total of 386 submissions at the different challenge stages and tracks, the challenge highlighted the difficulty of the tasks. While ten competing teams managed to outperform the challenge's baseline [37] on track 1 as detailed in Sec. 5, no team managed to surpass it on the track 2.

The paper summarizes the preparation and results of the UPAR-Challenge. In the following sections, we describe the challenge data preparation (Sec. 3), the challenge setup (Sec. 4.1), evaluation methodology (Sec. 4.2) and baseline (Sec. 4.3), a description of submitted methods (Sec. 5.1), the results (Sec. 5.3), and provide a brief discussion about future research directions (Sec. 6).

2. Related Work

Pedestrian Attribute Recognition: Pedestrian attribute recognition is a well-established research field. Deep learning-based research focuses on global models [15, 19, 21, 40, 50], part-based models [13, 22, 24, 46, 48, 52] aiming to partition the human body into several parts, or attention-based approaches [14, 18, 26, 32, 43]. Furthermore, recent works address co-occurrences of attributes in multiple ways. While Han et al. [15] argue that adjusting the attribute predictions based on co-occurrence priors is beneficial for the task, Zou et al. [51] demonstrate that especially generalization might suffer if unintended co-occurrence biases are learned. A currently emerging field is the use of transformers, as self-attention is capable of capturing long-range dependencies in contrast to convolutions. The PARFormer [11] is a transformer-based baseline model, whereas further approaches [4, 44] propose to encode the predictable attributes with transformers to capture inter-attribute correlations. Recent studies [5, 19, 37] indicate that thoroughly optimized global models achieve state-of-the-art performance, especially in terms of generalization to new data sources, as intricate models tend to overfit to the characteristics of biased research datasets.

Attribute-based Person Retrieval: Attribute-based person retrieval is typically tackled by either pedestrian attribute recognition or learning joint feature spaces between textual and image features. Methods utilizing attribute recognition compare discrete attribute queries with the attributes extracted for the images to be searched through [12, 21, 33–36,

[1] https://chalearnlap.cvc.uab.cat/challenge/57/description/

38, 39, 45]. In contrast, the latter approach extracts embeddings from both the query attributes as well as the images and aims at aligning them with different techniques. Some works [3,47] rely on generative approaches for the alignment of the two modalities. However, Jeong et al. argue that the training of such approaches is unstable. Thus, they propose novel loss and regularization functions to enable learning a robust embedding space with conventional training schemes. Similarly, Zhu et al. [53] introduce triplet loss-based loss functions that consider both inter-modal and intra-modal embedding differences. Further works [9, 17] extract and match features on multiple levels, i.e., global and local or attribute-specific embeddings.

3. MEVID Attributes

The Multi-view Extended Videos with Identities (MEVID) [7] is a dataset for large-scale, video person re-identification (ReID) in the wild. It is a subset of the MEVA dataset [6] which contains over 2,237 unique outfits, worn by 176 actors, each photographed front and back, with and without outerwear. This check in collection provide photos, each showing the actor holding a card which identifies their ID. Each check in photo is thus connected with a global ID which links each instance of a person in the MEVID dataset. The MEVID dataset itself contains 158 actors wearing 598 different outfits across 17 unique indoor and outdoor locations. The videos are collected from a wide range of natural scenes such as parking lots, bus stations, cafes, school environments, and more, for a total of 33 different camera viewpoints and a wide range of scales on each target, with the smallest on the order of 75 pixels and the largest of 500 pixels on target height. MEVID is divided into the train and test sets. The train set, contains 104 identities with 485 outfits in 6,338 tracklets. The test set includes 54 global identities with 113 outfits in 1,754 tracklets.

The MEVID dataset does not provide PAR annotation, which is the reason why we asked 3 paid annotators to manually define the 40 UPAR attributes for the check in collection. The annotation is performed for each person's outfit individually, which mean the attributes of the same person are described using the different images of this person in the same outfit (front and back). The annotation process follows the UPAR annotation process in [37]. Accordingly, for each frame, the body area is divided into the two areas, up the hips (Upper-body) and down the hips (Lower-body). The clothing for the upper-body and lower-body regions is classified by determining the color class for the UpperBodyClothingColor and LowerBodyClothingColor attributes, respectively. The color classes are divided into unique color classes (black, white, gray, red, blue, yellow, orange, green, purple, pink, purple), a collection class (other), and a class for missing assignments (unknown). The clothing in the lower body region is classified by determining the length class for the LowerBodyClothingLength attribute. The length classes can be divided into two unique classes (long, short) and one for missing assignments (unknown). If the person is not clearly underage (child, teenager) or clearly an adult or elderly, the attribute is set to unknown. To maintain consistency, each frame is checked according to the dual control principle by an annotator (validator) who has not previously classified the same frame. In this process, operating errors are detected and corrected directly by the validator. Please refer to supplementary material of [37] for more details. This results in the annotation of 2,848 images with 113,920 labels.

4. Challenge Design

The 2024 WACV Challenge for Pedestrian Attribute Recognition and Attribute-based Person Retrieval comprises two tracks linked to semantic descriptors of pedestrians, including clothing or gender information: PAR and attribute-based person retrieval. While both tracks use the same data sources, they differ in evaluation criteria. Unlike the prior competition, there exists only one training split, and data from multiple sources are permissible for training. Furthermore, validation is conducted using data from the same sources in order to replicate the realistic setting when actual target domain data is unavailable. The teams had the flexibility to conduct various experiments with subsets of the training and validation data or apply cross-validation schemes. Since the challenge aims to investigate methods that can generalize to new and potentially unknown domains without re-training, calibration, or domain adaptation, available information about the test set was limited. The private test set for this challenge comprises images from a single data source, featuring both indoor and outdoor images. In detail, the goals for each track are defined as follows:

- **Track 1: Pedestrian Attribute Recognition** - The objective of this track is the development of a PAR model that accurately recognizes the soft biometrics of individuals in the presence of domain shifts.
- **Track 2: Attribute-Based Person Retrieval** - The goal of this track is to locate individuals within a large gallery database who match a certain attribute description. Approaches should use binary attribute queries and gallery images as input, then rank the images based on their similarity to the query.

The challenge included two phases, development and test, during which participants were required to submit their results for a validation set using the public training data released during the development phase. During the final stage of the challenge, each participating team was permitted to submit only three entries, following the release of the test data a few days before the end of the challenge. Participants who surpassed the baseline and thus qualified for winning

Category	Age	Gender	Hair length	UB clothing length	UB clothing color	LB clothing length	LB clothing color	LB clothing type	Backpack	Bag	Glasses	Hat
Attributes	Young Adult Elderly	Female	Short Long Bald	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Trousers&Shorts Skirt&Dress	Backpack	Bag	Normal Sun	Hat

Table 1. **UPAR Attributes** – Attribute annotations included in the UPAR dataset. Source: [5]

the challenge were obligated to share their codes and trained models after the end of the challenge. This allows the organizers to replicate the results submitted during the test phase in a code verification stage.

4.1. UPAR-Challenge 2024 Dataset

Analogous to the first UPAR challenge[2] [5], the challenge utilizes an extension of the UPAR dataset[3] [5, 37]. The challenge dataset consists of the harmonization of three public datasets, namely PA100K [25], PETA [8], and Market1501-Attributes [23, 49], for training and validation, and a novel test set. 40 binary attributes have been unified between those for which we provide additional annotations. This dataset enables the investigation of PAR and attribute-based person retrieval methods' generalization ability under different attribute distributions, viewpoints, varying illumination, and low resolution. The novel test set builds on the image data included in the MEVID dataset [7].

Attribute annotations for this data are contributed, as described in Sec. 3. Note that annotations are provided on the identity level, *i.e.*, the attributes of each individual included in the data is annotated once rather than creating labels for each single image. Moreover, it is found that the dataset tracks contain a substantial amount of annotation errors, ranging from empty bounding boxes to incorrect assignment of identities and outfits. Thus, the dataset was manually curated and representative images are selected to form the test set and allow meaningful evaluation. In total, 28,095 test images were chosen. In conjunction with the training and validation splits, 159,201 person images are included in the UPAR-Challenge 2024 dataset, as reported in Table 3. The attribute annotations in this year's challenge are identical to those originally proposed for the UPAR dataset. An overview is given in Table 1.

4.2. Evaluation Protocol

The evaluation protocol differs from the UPAR challenge 2023. Data from multiple dataset domains is made available for training and validation. This procedure corresponds to the real-world scenario when diverse data is available for training but from data sources divergent to the target domain. Both challenge tracks leverage the same data splits. Only images specified for the training split were allowed for training. The use of any other data was strictly prohibited and verified during code verification. Table 2 illustrates the partitioning of UPAR sub-datasets to the splits. As the challenge aims at the examination of methods that generalize well to novel, unknown domains, any use of test data for re-training, calibration, or domain adaptions was prohibited. However, it was allowed to leverage the validation data during the test phase to train the final model.

The following metrics were considered to determine the challenge winners for the two tracks:

1. Harmonic mean from label-based mA and instance-based F1
2. mADM [36]

Training	Validation	Test
Market1501, PA100K, PETA	Market1501, PA100K, PETA	MEVID

Table 2. **UPAR Challenge 2024 domains** – Data from three datasets is used for training and validation, while images from the MEVID [7] dataset represent the target domain for generalization.

Split	# Images	# Queries
Training	97,699	–
Validation	33,407	3,462
Test	28,095	367

Table 3. **Split statistics** – The number of images used in both tracks, and the number of queries employed to evaluate track 2 are reported. In total, the UPAR-Challenge 2024 dataset consists of 159,201 person images.

[2] https://github.com/speckean/upar_challenge
[3] https://github.com/speckean/upar_dataset

The label-based mA measures the average between the positive and negative recall separately for the attributes, whereas the instance-based F1 score focuses on the predictions per instance, *i.e.*, how well the predicted attributes describe the depicted person. The mADM [36] is a new attribute-based person retrieval metric that takes into account the degree of agreement between the query attributes and ground truth labels of gallery samples, instead of only considering binary relevance. Consequently, produced quality scores better correspond to the visual perception of a retrieval ranking by humans and offer more detailed insights.

4.3. Baseline

The UPAR baseline proposed by Specker et al. [37] is employed as our challenge baseline due to its state-of-the-art performance in both challenge tasks. The model follows a straightforward classification architecture, comprising a ConvNeXt [28] backbone and a fully-connected classification head, thus, being computationally efficient. Similar to previous studies, PAR is treated as a multi-label classification problem with binary attributes. Consequently, the baseline incorporates a Sigmoid activation layer as the final layer and is trained using a weighted cross-entropy loss function [21]. The initial learning rate is set to 1e-4, and a plateau scheduler reduces it by a factor of 0.1 when the validation loss fails to decrease for more than four epochs. Weight decay is configured as 5e-4, and the AdamW optimizer [29] is utilized as it enhances generalization compared to the original Adam optimizer [20]. The baseline model is primarily designed for cross-domain attribute-based person retrieval, thus various techniques and modules are employed to prevent overfitting and improve generalization. These include exponential moving averages of model weights, appropriate batch sizes, label smoothing [41], dropout [31], and data augmentation techniques [16].

The implementation and training of the baseline are conducted using PyTorch 1.11 and CUDA 11.3 on NVIDIA GeForce RTX 3090 GPUs. Adaptive mixed precision is applied to accelerate the training process, and trainings are terminated once the validation accuracy ceases to improve.

5. Challenge Results

The challenge ran from 13 September 2023 to 28 October 2023 through Codabench[4], an open-source framework for running competitions. Track 1 of the challenge attracted a total of 60 registered participants. During the development phase, 21 active teams made a total of 296 submissions. Afterward, during the test phase, 32 active teams made a total of 72 submissions. The fewer submissions in the test phase come from the maximum number of submissions per participant in this final phase. It was set to 3 to prevent participants from improving their results by trial and error. Track 2 of the challenge attracted a total of 22 registered participants. During the development phase, only one active team made a total of 8 submissions. Afterward, during the test phase, five active teams made a total of 7 submissions. Since track 2 did attract much less attention and no approach was able to surpass the baseline, the results of this track are shortly reported and discussed.

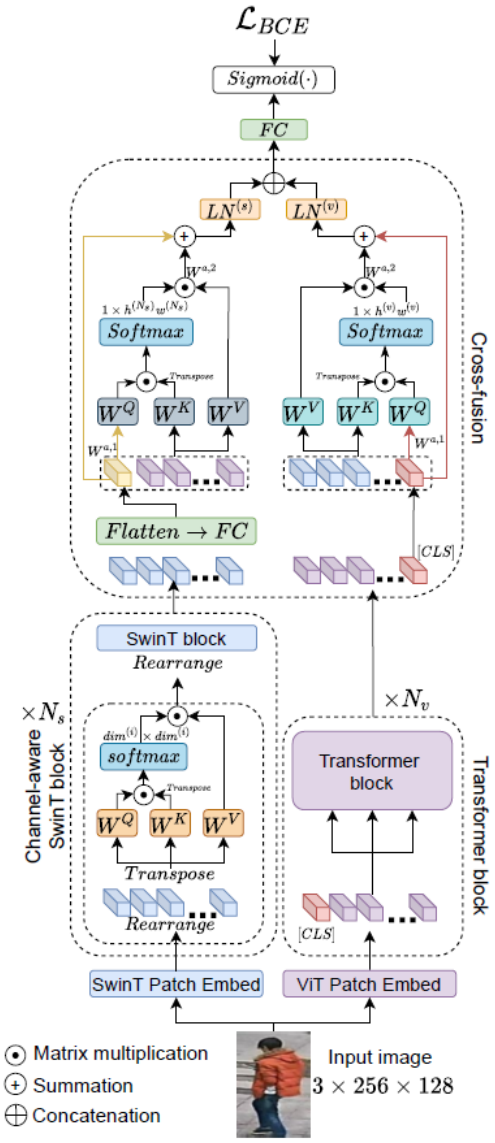

Figure 2. **Overview of the runner-up method of team *hdcolab* [1]** – Their overall architecture is composed of (i) two backbones for feature extraction: Eva vanilla ViT and a SwinT with channel-aware attention, and (ii) a cross-fusion module.

[4]https://www.codabench.org/

5.1. hdcolab

The runner-up team *hdcolab* proposed an extensive two-backbone architecture, illustrated in Fig. 2. Similarly to the winning solution of the precedent edition of this challenge, the authors chose a transformer architecture instead of a CNN as in the baseline. Their method called C²T-Net [1] uses two transformer networks, SwinT-base [27] and a customized variant of the vanilla vision transformer ViT-large [10] in order to capture both local and global aspects of a person. To this aim, they propose to customize each SwinT block with a channel-aware attention mechanism. At the end of the two branches, instead of concatenating the features vectors emerging from SwinT and ViT, the authors argue that this does not take fully advantage of the feature maps delivered by both backbones. Instead, they decide to use cross-attention to develop a cross-fusion module in order to support mutual benefits between local and global features. Finally, after fusion, a full-connected layer is used followed by a sigmoid activation. The whole model is trained using binary-cross entropy loss using RandomCrop and RamdomHorizontalFlip augmentation with an input resolution of 128×256 pixels.

The model is trained end-to-end with the Adam optimizer, a starting learning rate of 1e-6, a batch size of 64, and the plateau scheduler. This means that the learning rate is reduced by a factor of 0.1 when the results did not improve for 4 subsequent epochs.

5.2. DS

The method proposed by team *DS*[5] leveraged the CNN EfficientNetV2-Large backbone [42], attribute-wise balancing, and stochastic weight averaging to improve robustness and accuracy. For each attribute, a separate classification head consisting of two fully-connected layers with a ReLU and dropout layer in between are applied. Moreover, they design a variant of the cross-entropy loss with attribute-wise balancing for attribute-wise binary classification. It applies label smoothing [30] and weight balancing to the logits and labels batch-wise. To achieve this balance, they multiply the number of negative samples by the inverse of the batch size for positive samples, and vice versa for negative samples. Furthermore, they try to determine an optimal threshold for each attribute using the validation set to optimize the model's performance. To this aim, they use an attribute-based grid search approach similar to [35]. They initialize the threshold for each attribute at 0.5 and iteratively adjust the threshold for each attribute, seeking the best threshold that maximizes the mA metric. The search space for each attribute ranges from 0.1 to 0.9. Finally, the authors use the Albumentations Library [2] to perform exhaustive data augmentation using RGBShift, ColorJitter, ImageCompression, SafeRotate, Affine, HonrizontalFlip and CoarseDropout.

Training is conducted using the SGD optimizer with a learning rate of 0.01, weight decay of 1e-4, and momentum of 0.9. A batch size of 64 is employed, and the training is carried out for 40 epochs. The label smoothing parameter is set to 0.1 and stochastic weight averaging starts from epoch 10 with updates every 100 training iterations.

[5]Deeping Source Inc.: https://www.deepingsource.io/

Rank	Method	Avg.	mA	F1
1	fanttec	**71.8**	71.1	72.5
2	hdcolab†	71.7	70.5	**73.1**
3	sophere001	71.6	70.9	72.3
4	doanhbc	71.5	70.3	72.7
5	bhn2023	71.4	70.1	72.7
6	hungdoanhcnku	71.1	68.9	72.4
7	DS†	70.5	**72.2**	68.9
8	sidonio	70.0	71.7	68.4
9	thorin	70.0	71.1	68.9
10	hungda	69.8	72.1	67.7
-	UPAR Baseline [37]	69.4	68.0	70.9

Table 4. **Codabench leaderboard for track 1** – 10 teams managed to surpass the UPAR baseline concerning track 1. Best scores are highlighted in bold. †Approach that submitted code and a fact sheet.

Rank	Method	mADM	mAP	R-1
-	UPAR Baseline [37]	**43.1**	**13.4**	**26.2**
1	DS†	31.9	7.4	20.7
2	sidonio	30.4	6.7	16.1
3	thorin	30.4	6.8	16.1

Table 5. **Codabench leaderboard for track 2** – No approach managed to outperform the baseline. Best scores are highlighted in bold. †Approach that submitted code and a fact sheet.

5.3. Results & Findings

This section evaluates and discusses the challenge results for both tracks.

Track 1 - Pedestrian attribute recognition: Results for track 1 are given in Table 4. The leaderboard shows that, contrary to the first edition of the challenge, 10 teams managed to surpass the UPAR baseline in terms of track 1. This is attributed to the new evaluation protocol that allowed the use of diverse data from multiple image domains for training. As a result, training large architectures, as done by *hdcolab*, was feasible. The increased diversity in the training data reduces the risk of overfitting and learning biases that do not transfer well to new image sources. Therefore, more participants were able to outperform the simple and lightweight UPAR baseline. These findings demonstrate the importance of diverse training data in order to train large models and achieve

strong generalization performance, especially in finegrained classification tasks such as PAR. Furthermore, imbalances of approaches in terms of label-based mA and instance-based F1 score are evident. For instance, the *DS* team achieved the best mA results but performed worse than the UPAR baseline regarding instance-based F1. In contrast, *hdcolab*'s method leads to the best F1 performance, but to the detriment of mA when compared with other approaches. The reason is that depending on the model architecture, the focus is either on individual attributes (reflected by mA) or on entire descriptions obtained for persons. The multi-head structure of *DS* discards attribute correlations and applies separate classification heads for the attributes. Thus, the individual attributes are well recognized. Contrary to that, the model of *hdcolab* extracts global feature representations across all attributes for the person images, thereby implicitly leveraging correlations between attributes, such as co-occurrences. This approach results in better semantic descriptions of a person's attributes and, therefore, in enhanced instance-based F1 performance. In summary, the best method *fanttec* is capable of improving the challenge metric compared to the baseline by 2.4 points.

Track 2 - Attribute-based person retrieval: Analogous to the UPAR-Challenge 2023, none of the participants were able to outperform the UPAR baseline [37] concerning attribute-based person retrieval, as it is observable from the leaderboard results in Table 5. One possible explanation might be that the approaches still overfit the training data and, thus, get over- or underconfident regarding the recognition of the attributes. This negatively affects the confidence scores of PAR models, which are typically employed to calculate the distance to the query to build the retrieval ranking [35, 37]. In contrast, PAR metrics are computed based on binarized attribute predictions and, hence, are less influenced by this issue. In conclusion, the simple and lightweight UPAR baseline still represents the state-of-the-art concerning attribute-based person retrieval in generalization settings by a large margin. In terms of mADM, the difference to the second-best approach by *DS* are a remarkable 11.2 points.

6. Conclusions

The UPAR-Challenge 2024 attracted over 82 participants, who made 304 submissions during validation and 79 submissions for the test set. This second edition of the challenge, introduced a new evaluation set based on the MEVID dataset, for which 1.1M additional binary annotations were contributed. This time, 10 teams managed to surpass the UPAR baseline for track 1, which is related to the new evaluation protocol that allowed the use of diverse data from multiple image domains for training. In track 2, for which a new evaluation metric was introduced, participants could not manage to beat the proposed baseline. The challenge and its results highlight the difficulty of PAR and attribute-based person retrieval in real-world surveillance scenarios. While the overall winner of track 1 proposed a balanced method in terms of label-based mA and instance-based F1 score, the solution with the best performance concerning the mA metric is still largely surpassed by the baseline w.r.t. to the instance-based F1 metric. The runner-up solution surpassed the winner for F1, but could not compete for mA. This shows that the evaluation protocols designed for real-world challenges were difficult, and that the PAR task remains an open field for future research. Following the release of the UPAR dataset [37], the UPAR-Challenge 2023 [5] and these results, we see a positive development towards research on PAR with single images for realistic application scenarios. However, we could not observe this trend for the closely related task of attribute-based person retrieval. With upcoming regulation for AI and data-privacy, we emphasize that future research should focus more closely on this task, which shows favorable privacy-preserving properties through the use of soft-biometric attributes. In future works we will focus on video PAR, as in real-world surveillance scenarios typically videos or tracks of persons are available rather than single images. This enables more robust recognition of individuals' semantic characteristics, since richer information about the appearance is available.

Acknowledgments

This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia program.

References

[1] Doanh C. Bui, Thinh V. Le, and Ba Hung Ngo. C^2T-net: Channel-aware cross-fused transformer-style networks for pedestrian attribute recognition. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2024. 5, 6

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: fast and flexible image augmentations. 11(2):125, 2020. 6

[3] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12359 of *Springer eBook Collection*, pages 230–247. Springer International Publishing and Imprint Springer, 1st ed. 2020 edition, 2020. 3

[4] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6994–7004, 2022. 2

[5] Mickael Cormier, Andreas Specker, Julio C. S. Jacques, Lucas Florin, Jurgen Metzler, Thomas B. Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jurgen Beyerer. Upar challenge:

Pedestrian attribute recognition and attribute-based person retrieval - dataset, design, and results. In *2023 IEEE Winter Conference on Applications of Computer Vision workshops*, pages 166–175. Institute of Electrical and Electronics Engineers and Computer Vision Foundation, IEEE, 2023. 2, 4, 7

[6] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021. 3

[7] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, and Brian Clipp. Mevid: Multi-view extended videos with identities for video person re-identification. In *2023 IEEE Winter Conference on Applications of Computer Vision*, pages 1634–1643. IEEE Computer Society and Computer Vision Foundation, IEEE, 2023. 1, 2, 3, 4

[8] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In Kien A. Hua, editor, *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792. ACM, 2014. 1, 2, 4

[9] Qi Dong, Xiatian Zhu, and Shaogang Gong. Person search by text attribute query as zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv, 2021. 6

[11] Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, page 1, 2023. 2

[12] Lucas Florin, Andreas Specker, Arne Schumann, and Jürgen Beyerer. Hardness prediction for more reliable attribute-based person re-identification. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 418–424. Institute of Electrical and Electronics Engineers (IEEE), 2021. 2

[13] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *2015 IEEE International Conference on Computer Vision*, pages 2470–2478, Piscataway, NJ, 2015. IEEE. 2

[14] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. 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) // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–739. IEEE, 2019. 2

[15] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition. In *IJCAI*, 2019. 2

[16] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 5

[17] Sara Iodice and Krystian Mikolajczyk. Text attribute aggregation and visual feature decomposition for person search. In *BMVC*, 2020. 3

[18] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In Eric Mortensen, editor, *2021 IEEE/CVF International Conference on Computer Vision*, pages 942–951. Institute of Electrical and Electronics Engineers (IEEE) and Computer Vision Foundation, IEEE, 2021. 2

[19] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Third IAPR Asian Conference on Pattern Recognition - ACPR 2015*, pages 111–115. International Association for Pattern Recognition, IEEE, 2015. 2, 5

[22] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 684–700. Springer, 2016. 2

[23] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. 95:151–161, 2019. Accepted to Pattern Recognition (PR). 2, 4

[24] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. 2018. 2

[25] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017. 4

[26] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *2017 IEEE International Conference on Computer Vision*, IEEE Xplore Digital Library, pages 350–359. Institute of Electrical and Electronics Engineers (IEEE), IEEE, 2017. 1, 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 6

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 5

[30] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? 32, 2019. 6

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. 15(56):1929–1958, 2014. 5

[32] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision*, pages 680–697, 2018. 2

[33] Walter J Scheirer, Neeraj Kumar, Peter N Belhumeur, and Terrance E Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[34] Arne Schumann, Andreas Specker, and Jürgen Beyerer. Attribute-based person retrieval and search in video sequences. IEEE, 2018. 2

[35] Andreas Specker and Jürgen Beyerer. Improving attribute-based person retrieval by using a calibrated, weighted, and distribution-based distance metric. In *2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, Anchorage, AK, USA, 19-22 Sept. 2021*, pages 2378–2382. Institute of Electrical and Electronics Engineers (IEEE), 2021. 2, 6, 7

[36] Andreas Specker and Jürgen Beyerer. Balanced pedestrian attribute recognition for improved attribute-based person retrieval. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE, 7/4/2023 - 7/7/2023. 2, 4, 5

[37] Andreas Specker, Mickael Cormier, and Jurgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *2023 IEEE Winter Conference on Applications of Computer Vision*, pages 981–990. IEEE Computer Society and Computer Vision Foundation, IEEE, 2023. 2, 3, 4, 5, 6, 7

[38] Andreas Specker, Arne Schumann, and Jürgen Beyerer. An interactive framework for cross-modal attribute-based person retrieval. In *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, Taipei, Taiwan, 18-21 Sept. 2019*, page 8909832. Institute of Electrical and Electronics Engineers (IEEE), 2019. 2

[39] Andreas Specker, Arne Schumann, and Jurgen Beyerer. An evaluation of design choices for pedestrian attribute recognition in video. In *2020 IEEE International Conference on Image Processing*, pages 2331–2335. Institute of Electrical and Electronics Engineers and IEEE Signal Processing Society, IEEE, 2020. 2

[40] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015. 2

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826. IEEE, IEEE, 2016. 5

[42] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 6

[43] Chufeng Tang, Lu Sheng, Zhao-Xiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[44] Zengming Tang and Jun Huang. Drformer: Learning dual relations using transformer for pedestrian attribute recognition. 497:159–169, 2022. PII: S0925231222005598. 2

[45] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, 2009. 2

[46] Luwei Yang, Ligeng Zhu, Yichen Wei, Shuang Liang, and Ping Tan. Attribute recognition from adaptive parts. In R. C. Wilson, E. R. Hancock, W. A. P. Smith, N. E. Pears, and A. G. Bors, editors, *Procedings of the British Machine Vision Conference 2016*, pages 81.1–81.11. British Machine Vision Association, 2016. 2

[47] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. International Joint Conferences on Artificial Intelligence Organization, 2018. 3

[48] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. IEEE, 2014. 2

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision*, pages 1116–1124. IEEE International Conference on Computer Vision and Institute of Electrical and Electronics Engineers (IEEE), IEEE, 2015. 1, 2, 4

[50] Jiabao Zhong, Hezhe Qiao, Lin Chen, Mingsheng Shang, and Qun Liu. Improving pedestrian attribute recognition with multi-scale spatial calibration. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[51] Yibo Zhou, Hai-Miao Hu, Jinzuo Yu, Zhenbo Xu, Weiqing Lu, and Yuran Cao. A solution to co-occurence bias: Attributes disentanglement via mutual information minimization for pedestrian attribute recognition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2023. 2

[52] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *2015 International Conference on Biometrics (ICB)*. IEEE, 2015. 2

[53] Jianqing Zhu, Liu Liu, Yibing Zhan, Xiaobin Zhu, Huanqiang Zeng, and Dacheng Tao. Attribute-image person re-identification via modal-consistent metric learning. *International Journal of Computer Vision*, 131(11):2959–2976, 2023. 3