# UPAR Challenge: 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[4],
Lucas Florin[2], Jürgen Metzler[2,3], Thomas B. Moeslund [5],
Kamal Nasrollahi[5,6], Sergio Escalera[4,5,7], Jürgen Beyerer[2,1,3]

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

## Abstract

*In civilian video security monitoring, retrieving and tracking a person of interest often rely on witness testimony and their appearance description. Deployed systems rely on a large amount of annotated training data and are expected to show consistent performance in diverse areas and generalize well between diverse settings w.r.t. different viewpoints, illumination, resolution, occlusions, and poses for indoor and outdoor scenes. However, for such generalization, the system would require a large amount of various annotated data for training and evaluation. The WACV 2023 Pedestrian Attribute Recognition and Attributed-based Person Retrieval Challenge (UPAR-Challenge) aimed to spotlight the problem of domain gaps in a real-world surveillance context and highlight the challenges and limitations of existing methods. The UPAR dataset, composed of 40 important binary attributes over 12 attribute categories across four datasets, was extended with data captured from a low-flying UAV from the P-DESTRE dataset. To this aim, 0.6M additional annotations were manually labeled and validated. Each track evaluated the robustness of the competing methods to domain shifts by training on limited data from a specific domain and evaluating using data from unseen domains. The challenge attracted 41 registered participants, but only one team managed to outperform the baseline on one track, emphasizing the task's difficulty. This work describes the challenge design, the adopted dataset, obtained results, as well as future directions on the topic.*

## 1. Introduction

Person Attribute Recognition (PAR) and attribute-based person retrieval in surveillance data are challenging tasks on single domains due to limited image quality, strongly localized attributes, and limited visibility due to varying viewing angles or occlusions. PAR aims at recognizing persons' semantic attributes, such as gender, age, or information about clothing. Attribute-based retrieval systems may build on PAR methods and allow searching through an extensive database of person images for individuals matching a specific set of semantic attributes. For deployment and long-term use of such machine learning algorithms in a surveillance context, the algorithms must be robust to domain gaps that occur when the environment changes.

The domain gap between five PAR datasets is illustrated in Fig. 1 The Market1501 [38, 16] dataset has mostly images with low resolution. Also, many individuals are interacting with objects, such as riding a bicycle or carrying things. The images are closely cropped, sometimes too close, such as in the rightmost image, which results in only partly visible persons. The PA100K [18] dataset has images with higher resolution. However, there are several cropping errors where the legs, head, or both are missing. The images in the P-DESTRE [10] dataset were captured using UAVs. Thus, the images show steep camera angles in comparison with datasets recorded by static cameras. Furthermore, this dataset contains several cropping errors where only the head and torso are visible. The PETA [2] dataset consists of low-resolution images with a different cropping scheme than the other datasets. Large amounts of images have an extensive

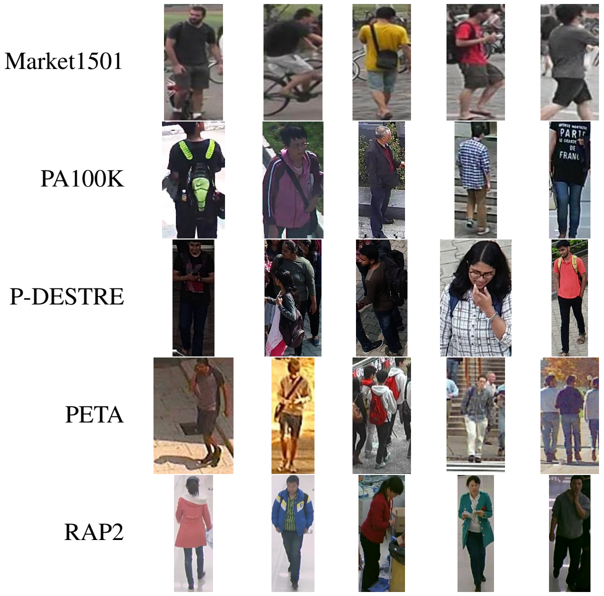

Figure 1: **PAR datasets –** Sample images from the five datasets composing the UPAR-Challenge dataset. Each dataset shows different characteristics and thus poses different challenges.

crop, causing images to contain multiple individuals. Besides, lighting conditions vary significantly across images. These datasets are mainly captured in an outdoor environment. The RAP2 [15] dataset, on the other hand, is captured indoors in a shopping mall. Due to this, most pictures have better lighting. In general, the main differences in the datasets come from different resolutions, camera angles, and cropping.

The WACV 2023 Pedestrian Attribute Recognition and Attributed-based Person Retrieval Challenge (UPAR-Challenge)[1] aims to demonstrate the problem of domain gaps in a real-world surveillance context and highlight the challenges and limitations of existing methods to provide a direction for future research. 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. To inspect the consistency of performance across varied scenarios, the UPAR Dataset [27], composed of four of the aforementioned datasets PA100K, PETA, RAP2, and Market1501, is extended with the P-DESTRE dataset to increase diversity further. To this aim, a total of 0.6M new binary annotations are contributed to P-DESTRE.

The UPAR Challenge 2023 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 limited data from a specific domain and evaluating using data from multiple unseen domains. The challenge attracted a total of 41 registered participants on its different tracks. With a total of 94 submissions at the different challenge stages and tracks, the challenge highlighted the difficulty of the task. Only one solution managed to outperform the challenge's baseline [27] on track 1 as detailed in Sec. 4.1.

The paper summarizes the preparation and results of the UPAR-Challenge. In the following sections, we describe the challenge setup (Sec. 3.3), challenge data preparation (Sec. 3.1), evaluation methodology (Sec. 3.2), analysis of submitted results (Sec. 4.2), and a brief discussion of insights and future research directions (Sec. 4.3).

## 2. Related Work

**Pedestrian Attribute Recognition**. The first deep-learning-based approaches to pedestrian attribute recognition assumed the task to be a multi-label classification task, training models with cross-entropy loss. To counterbalance biases in attribute distributions in the datasets DeepMAR [11] gives higher weights to rare attributes during training. Recently, approaches such as attention mechanisms for focusing on the correct regions of an image [18, 17, 13, 35, 6, 23], and using multi-scale features [37, 23, 31, 34, 39] have achieved gains in performance over earlier works. However, these approaches tend to have highly complex architectures. Recent publications have shown that it is possible to get state-of-the-art performance using only a backbone model and simple tricks to improve training [8, 27] or applying spatial and semantic regularization [7].

The above mentioned approaches achieve strong performance on single datasets [16, 18, 2, 11, 14] that usually focus on only one scenario. The UPAR dataset [27] unifies four different datasets on a common set of labeled attributes and allows studying how well these models generalize over the domain gap that exists between different scenarios. In this work, we contribute annotations for the P-DESTRE dataset to UPAR in order to increase diversity w.r.t. camera angles and image resolutions.

**Attribute-based Person Retrieval**. The first kind of approaches of solving attribute-based person retrieval is by training a PAR model. Then, given an attribute description as a query, the system retrieves images that most closely match the given attribute vector, making the result explainable [32, 24, 12, 25, 29, 26, 28, 5]. The other approach is to align attribute descriptions and image embeddings in a

[1] https://chalearnlap.cvc.uab.cat/challenge/52/description/

shared cross-modal feature space. This can be implemented by, *e.g.*, using high-dimensional hierarchical embeddings and an additional matching network [3] or by matching person attributes and images in a joint feature space [36, 1].

## 3. Challenge Design

The WACV 2023 Pedestrian Attribute Recognition and Attribute-based Person Retrieval Challenge is split into two tracks associated with semantic pedestrian attributes, such as gender or clothing information: PAR and attribute-based person retrieval. Both tracks were built on the same data sources, but had different evaluation criteria. Three different dataset splits for both tracks use other training domains. Each track evaluates how robust a given method is to domain shifts by training on limited data from a specific domain and evaluating using data from several out-of-distribution domains. Both tasks use the same image data for training and evaluation. In detail, the tasks to solve in the tracks are defined as follows:

- **Track 1: Pedestrian Attribute Recognition:** The task in this track is to develop and train an attribute classifier that accurately predicts persons' semantic attributes under domain shifts.

- **Track 2: Attribute-based Person Retrieval:** Attribute-based person retrieval aims to find persons in a vast gallery database that match a specific attribute description. Approaches should take binary attribute queries and gallery images as input and rank the photos according to their similarity to the query.

Each track was composed of two phases, *i.e.*, the development and test phases. During the development phase, the public training data was released, and participants were required to submit their predictions concerning a validation set. During the subsequent test phase, participants needed to submit their results for the test data, which was released just a few days before the end of the challenge. Participants who beat the baseline and are thus candidates for winning the challenge were required to share their codes and trained models after the end of the challenge so that the organizers could reproduce the results submitted at the test phase in a *code verification stage*. At the end of the challenge, top-ranked methods that passed the code verification stage were considered valid submissions and were applied to a private test dataset for final ranking.

### 3.1. UPAR-Challenge Dataset

The challenge dataset[2] used in the UPAR Challenge at WACV'23 is an extension of the existing UPAR dataset [27][3], which provides annotations for 40 binary attributes over 12 categories from four datasets (detailed list in Tab. 1). This dataset enables the investigation of attribute recognition and attribute-based person retrieval methods' generalization ability under different attribute distributions, viewpoints, varying illumination, and low resolutions. The public part of the challenge dataset consists of the harmonization of three public datasets (PA100K [18] (outdoor), PETA [2] (mixed), and Market1501-Attributes [16, 38] (outdoor)). The private part of the challenge test dataset is composed of RAP2 [14] (indoor) and P-DESTRE [10] (outdoor). The P-DESTRE dataset was recorded with drones flying between 5.5 and 6.7 meters in height over different scenes of the campuses of two universities in Portugal and India. We asked 16 paid annotators to manually define the color of the upper-body and lower-body clothing for eleven unique colors plus additional classes to indicate multiple colors or colors not included in the color list. Furthermore, annotations for the lower-body and upper-body clothing lengths were assigned in a further iteration. Thus, we provide 0.6M manually labeled and validated new binary annotations for 22,518 images for the P-DESTRE dataset.

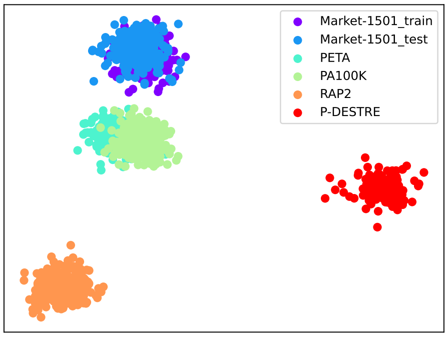

Figure 2: **UPAR-Challenge Domains** – Distribution of image embeddings extracted using an ImageNet pre-trained Inception model and projected into two-dimensional features using linear discriminant analysis. Train and test set embeddings from the same dataset overlap entirely. In contrast, the UPAR-Challenge dataset, as a combination of multiple sub-datasets with disjunct data distributions, poses a more realistic and challenging problem and requires models to generalize well across different domains.

As can be seen in Fig. 2, the distributions of the data vary significantly compared to different sets of the same dataset. For instance, for the Market1501-dataset the train and test set embeddings from the same dataset overlap entirely. In contrast, embeddings (produced by an Inception model trained on ImageNet) of the RAP2 and P-DESTRE show apparent disjunct data distributions, which poses a

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

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

| 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 | Female | Short | Short | Black | Short | Black | Trousers&Shorts | Backpack | Bag | Normal | Hat |
| | Adult | | Long | | Blue | | Blue | Skirt&Dress | | | Sun | |
| | Elderly | | Bald | | Brown | | Brown | | | | | |
| | | | | | Green | | Green | | | | | |
| | | | | | Grey | | Grey | | | | | |
| | | | | | Orange | | Orange | | | | | |
| | | | | | Pink | | Pink | | | | | |
| | | | | | Purple | | Purple | | | | | |
| | | | | | Red | | Red | | | | | |
| | | | | | White | | White | | | | | |
| | | | | | Yellow | | Yellow | | | | | |
| | | | | | Other | | Other | | | | | |

Table 1: **UPAR Attributes** – Attribute annotations included in the UPAR dataset.

| Split ID | Training | Public Evaluation | Private Evaluation |
|---|---|---|---|
| 0 | Market1501 | PA100K, PETA | RAP2, P-DESTRE |
| 1 | PA100K | Market1501, PETA | RAP2, P-DESTRE |
| 2 | PETA | Market1501, PA100K | RAP2, P-DESTRE |

Table 2: **UPAR-Challenge splits** – The challenge uses a Cross-validation (CV) protocol, *i.e*., there are three splits of training and validation data. Only data from one domain is used for training in each split, so that evaluation is performed on unseen domains.

| Split ID | # Training images | # Attributes with positive sample |
|---|---|---|
| 0 | 10,000 | 35 |
| 1 | 79,001 | 40 |
| 2 | 8,668 | 39 |

Table 3: **Number of training images** – Number of training images per split and the number of attributes that are used for evaluation, *i.e*., attributes with at least one training sample. Validation and public testing for track 1 is done using 14,580 and 33,407 images, respectively.

| Split ID | # Val queries | # Val gallery images | # Test queries | # Test gallery images |
|---|---|---|---|---|
| 0 | 2,267 | 11,656 | 2,706 | 16,949 |
| 1 | 1,325 | 4,658 | 2,557 | 23,421 |
| 2 | 2,336 | 12,846 | 2,169 | 26,444 |

Table 4: **UPAR-Challenge public evaluation splits** – The number of queries and gallery images for track 2.

more realistic and challenging problem and requires models to generalize well across the different domains. Following [27], this challenge applies a cross-domain evaluation scheme using three different splits to assess the generalization performance of the submitted methods for the public evaluation, as described in Tab. 2. Statistics for track 1 and track 2 are provided in Tab. 3 and Tab. 4 respectively.

### 3.2. Evaluation Protocol

The challenge used a cross-validation evaluation protocol, *i.e*., there were three splits for training, validation, and test data. As the task was to develop models that generalize well to other domains, only data from one domain is used for training in each split. Evaluation was performed on image data originating from several domains. Since the models should be applicable to multiple unseen domains without any changes, it was not allowed to use different models, hyper-parameters, or approaches for different sub-sets of evaluation data within the same split. The training data was identical for both tracks, and training splits are defined as illustrated in Tab. 2.

Only images specified for the train split were allowed for training. The use of any other data was strictly prohibited and checked during code verification. Some attributes did not have positive samples in the training data of a split. Such attributes were ignored during the evaluation to receive meaningful results.

Since the challenge aimed to investigate methods that generalize well on new and possibly unknown domains without re-training, calibration, or domain adaptation, we only provided little information about the private test set. The final challenge winners were selected based on the score achieved on the evaluation server and the performance on the private test set.

Different evaluation metrics are used for the two tracks:

1. Harmonic mean from mA and instance-based F1

2. Harmonic mean from mAP and R-1

Since for rare attributes, approaches may achieve high accuracy by consistently predicting the absence of an attribute, the label-based mean accuracy (mA) and the instance-based F1 score are reported. The difference between mA and instance-based F1 score is that the mA calculation considers each attribute separately, while the instance-based F1 score measures the quality of predictions of all attributes with respect to the persons. First, the metrics were computed separately for each of the three splits and then averaged across the splits.

### 3.3. Baseline

We use the UPAR baseline proposed by Specker et al. [27] as our challenge baseline since it achieves state-of-the-art results for both challenge tasks. The model follows a straightforward classification architecture consisting of a ConvNeXt [20] backbone and a fully-connected classification head. Similar to related works, PAR is considered a multi-label classification problem with binary attributes. Therefore, the baseline includes a Sigmoid activation layer as the final layer and is trained with a weighted cross-entropy loss function [11]. The learning rate is initialized with $1e-4$, and a plateau scheduler lowers it by a factor of 0.1 once the validation loss has not decreased for more than four epochs. Weight decay was set to $5e-4$, and the AdamW optimizer [21] was applied since it improves generalization ability compared to the vanilla Adam [9]. The baseline model was primarily developed with the focus on cross-domain attribute-based retrieval. So, multiple tricks and modules are applied to avoid overfitting and improve generalization. For example, exponential moving averages of model weights, suitable batch sizes, label smoothing, dropout, and data augmentation techniques are leveraged.

The baseline is implemented and trained using PyTorch 1.11 and CUDA 11.3 on NVIDIA GeForce RTX 3090 GPUs. To speed up the training processes, adaptive mixed precision is applied and trainings are aborted as soon as the validation accuracy stops to improve.

## 4. Challenge Results

The challenge ran from 15 September 2022 to 31 October 2022 through Codalab[4][22], an open-source framework for running competitions. Track 1 of the challenge attracted a total of 30 registered participants. During the development phase, two active teams made a total of 67 submissions. We assume most teams chose to use the training data for cross-validation offline rather than the public development leaderboard. Afterward, during the test phase, six active teams made a total of 9 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. Furthermore, since several teams could not surpass the baseline in the development phase, we believe those teams needed more time to improve their submission for the test phase. Since track 2 did not attract much attention and no approach was able to surpass the baseline, the results of this track are shortly reported and discussed.

| Rank | Method | Challenge Avg. | mA | F1 |
|---|---|---|---|---|
| 1 | melaeric | **76.8** | **75.8**±1.6 | 77.9±2.5 |
| - | UPAR Baseline [27] | 75.3 | 71.5±1.9 | **79.6**±3.0 |
| 2 | Jai_C21 | 72.6 | 70.2±2.9 | 75.2±2.7 |
| 3 | harshtripathi6 | 72.1 | 66.5±1.8 | 78.8±1.9 |
| 4 | jzsherlock | 71.3 | 67.6±1.5 | 75.5±3.3 |
| 5 | ko4ro | 69.3 | 66.6±2.9 | 72.3±6.1 |
| - | Strong Baseline [8] | 69.3 | 66.2±3.0 | 72.6±5.3 |
| 6 | neptune | 65.2 | 62.8±2.9 | 67.8±3.3 |

Table 5: **Codalab Leaderboard for track 1 on the public test set** – Best scores are highlighted in bold. Only one team managed to surpass the challenge baseline [27]. The Strong Baseline [8] is included as a supplementary baseline and not ranked.

| Rank | Method | Challenge Avg. | mA | F1 |
|---|---|---|---|---|
| 1 | melaeric | **78.0** | **75.4**±3.7 | 80.9±2.7 |
| - | UPAR Baseline [27] | 76.0 | 71.0±2.8 | **81.7**±2.4 |

Table 6: **Leaderboard for track 1 on the private test set** – Best scores are highlighted in bold. Similar to the public leaderboard, *melaeric* outperforms the baseline w.r.t. mA and the challenge average score.

### 4.1. Leaderboard

Both tracks' public leaderboards at the test phase are shown in Tab. 5 and Tab. 7. We include the Strong Baseline [8] and the UPAR Challenge baseline in the ranking for orientation in track 1. All teams except one are able to surpass the Strong Baseline, which was not specially designed for cross-domain PAR. Only one team, *melaeric* [30], achieved better mA than the UPAR baseline, while none achieved a greater F1 score. This team achieved the best mA score by a large margin of 4.3 points against the UPAR Challenge baseline. In contrast, their results in terms of F1 score were worst than the UPAR baseline by 1.7 points. This team leads on the overall ranking with a significant margin of 1.5 points in challenge average against the UPAR Challenge baseline and 4.2 points against the next participant. After code verification, the top-1 solution is evaluated against the private test set. As shown in Tab. 6, the winning solution surpasses the baseline again with a large margin for the challenge average. While the standard deviation of team *melaeric*'s method grows larger, the mA results outperform the baseline by 4.4 points. Again, the baseline achieves a higher F1 by 0.8 points.

Regarding track 2, the baseline performs significantly better than *melaeric*. A possible explanation is the focus of their solution to improving mA at the cost of the F1-score, which, as an instance-based metric, is a more reliable predictor of retrieval performance.

[4] https://codalab.lisn.upsaclay.fr/

| Rank | Method | Challenge Avg. | mAP | R-1 |
|---|---|---|---|---|
| - | UPAR Baseline [27] | **14.0** | **12.4** | **16.1** |
| 1 | melaeric | 10.8 | 9.4 | 12.6 |

Table 7: **Codalab Leaderboard for track 2 on the public test set** – Best scores are highlighted in bold. No approaches managed to outperform the baseline.

Next, we briefly introduce the winning solution that passed the code verification stage based on the information provided by the authors.

### 4.2. Top-1 Team melaeric

The winning team *melaeric* proposed an extensive network architecture illustrated in Fig. 3. While the UPAR Challenge is based on a single-branched CNN baseline, this method bases on a more complex three-branched architecture with various and diverse elements. They used a Swin-Base [19] as their Transformer [4] backbone. The extracted features are then forwarded to an Attribute and Contextual Feature Projection module, where spatial contextual and attribute-specific features are learned individually. To this aim, this module is composed of two parallel parts: a spatial projection and an attribute projection. As shown in Fig. 4, different receptive fields with kernel sizes of 3×3, 5×5, and 7×7 are adopted in three parallel branches and subsequently concatenated to capture spatial contextual information. The attribute projection part builds attribute features by matrix multiplication between modulated features and attention masks. These attribute and contextual features are then concatenated in their Relation Exploration Module, which consists of a Graph Convolutional Network [33] and a Transformer to capture relationships among spatial features and attribute-specific features. Their training is formulated as a multi-label classification problem and adopts binary cross-entropy loss to supervise their training at three places. Considering the distribution imbalance between attributes, they introduced a penalty term based on the positive sample ratio of each attribute in the training set. In terms of data augmentation, random horizontal flipping, cropping, scaling, translation, erasing [40], and gaussian blur are applied to the images during training to reduce overfitting. This approach is similar but less extensive than the augmentation strategy used in the UPAR Challenge baseline.

Finally, a noticeable difference to the UPAR Challenge baseline is that this model dealt with image resized to $384 \times 384$ Pixels instead of $192 \times 256$ pixels.

### 4.3. Results & Findings

In this section, we compare the UPAR baseline [27] to the top-1 approach for track 1 of the challenge and discuss findings concerning PAR across multiple image domains with different characteristics and attribute distributions.

**Training data**. First, we investigate the influence of the training data on the generalization performance. Most approaches on the leaderboard, including the baseline and the winning approach, achieve the best results for split 1 due to the largest amount of training data. This finding is also valid for the private test set, as shown in Tab. 8. Since the split contains about eight times the number of training images than the other ones, more diverse and realistic scenarios with respect to, *e.g.*, attribute distributions are reflected. In addition, more positive samples for the single attributes allow more robust recognition of rarely occurring attributes, such as seldom colors.

**Evaluation data**. Next, we examine the results of the different evaluation datasets. Regarding mA, the best scores are achieved on the P-DESTRE dataset. The reason is the high resolution of the images. Even small characteristics, such as glasses, are clearly visible and can thus be better recognized by the models. Worst mA results were observed for the RAP2 dataset regardless of the training data used. Deeper investigations indicate that especially color attributes generalize poorly on the RAP2 dataset and that imbalanced attribute distribution severely impacts the results. Concerning the instance-based F1, the highest scores are achieved for RAP2, P-DESTRE, and Market1501, while especially on PETA the F1 is significantly lower.

**Label-based vs. instance-based metrics**. The results in Tab. 8 clearly show that *melaeric*'s approach leads to consistently higher label-based mA scores while the baseline is advantageous if the focus is on the instance-based F1. It is expected that approaches are biased toward either label- or instance-based metrics dependent on the method's design. *Melaeric* applies Transformers and an Attention module to improve the localization and recognition of fine-grained attributes. As a result, melaeric's method is better suited to recognize, *e.g.*, the glasses attributes compared to the UPAR baseline. In contrast, the baseline was developed with the focus on attribute-based person retrieval and, therefore, instance-based metrics. To achieve promising retrieval performance, it is more important to get consistent attribute predictions with respect to the entire appearance of persons rather than recognizing every occurrence of individual, highly localized attributes. This leads to the fact that the UPAR baseline clearly outperforms the track 1 winner on track 2, *i.e.*, the retrieval task since instance-based scores are better.

**Image domain shifts**. Fig. 2 visualizes the different distributions of the sub-dataset images. It can be seen that *e.g.*, the RAP2 and P-DESTRE data can be clearly distinguished

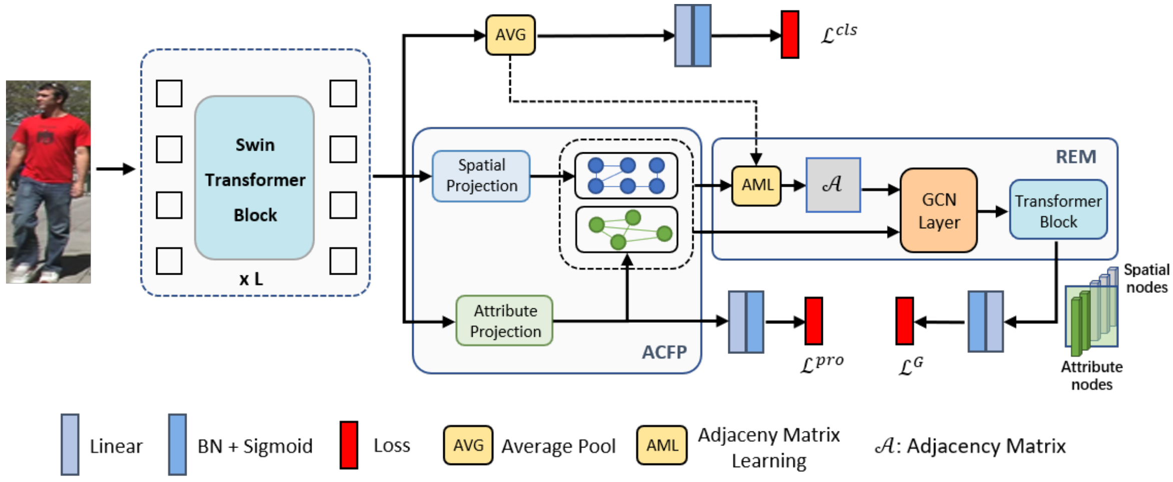

Figure 3: **Overview of the winning solution of team *melaeric*** – Their overall architecture is composed of (i) Feature Extraction, (ii) Attribute and Contextual Feature Projection (ACFP) and (iii) Relation Exploration Module (REM).

| Approach | Split | Average | | Market1501 | | PA100K | | PETA | | RAP2 | | P-DESTRE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mA | F1 | mA | F1 | mA | F1 | mA | F1 | mA | F1 | mA | F1 |
| melaeric | 0 | 72.6 | 75.8 | – | – | 76.8 | 77.9 | 70.7 | 71.0 | 69.0 | 77.8 | 73.7 | 76.6 |
| | 1 | **79.0** | 81.9 | **76.6** | 83.1 | – | – | **78.7** | 77.3 | **76.6** | 83.1 | **84.1** | 84.1 |
| | 2 | 75.3 | 80.5 | 73.4 | 79.1 | **78.6** | 79.0 | – | – | 72.3 | 83.4 | 76.7 | 80.3 |
| | Avg | 75.6 | 79.1 | 75.0 | 81.1 | 77.7 | 78.5 | 74.7 | 74.2 | 72.6 | 81.4 | 78.2 | 80.3 |
| UPAR Baseline [27] | 0 | 69.2 | 77.2 | – | – | 73.0 | 78.7 | 67.1 | 72.4 | 66.2 | 78.7 | 70.6 | 78.9 |
| | 1 | 74.6 | **83.7** | 74.9 | **85.9** | – | – | 73.4 | **79.5** | 72.0 | **84.1** | 78.0 | **85.2** |
| | 2 | 70.0 | 81.1 | 69.8 | 81.2 | 70.9 | **79.6** | – | – | 68.6 | 83.5 | 70.8 | 80.2 |
| | Avg | 71.3 | 80.4 | 72.4 | 83.6 | 72.0 | 79.2 | 70.3 | 76.0 | 68.9 | 82.1 | 73.1 | 81.4 |

Table 8: **Detailed results** – Comparison of the baseline with the winning approach. Best results are highlighted in bold. While *melaeric*'s method achieves better mA values, the baseline achieves higher F1 scores. Regarding splits, best scores are achieved when training on split 1, which has the most training data.

based on general image features. However, the challenge results indicate that these distribution shifts regarding image data only have a minor influence on the generalization performance. For instance, training the winning approach on split 1 and comparing the results on PETA and P-DESTRE, P-DESTRE scores with respect to both metrics are much higher, although the more significant domain gap concerning image data. Being robust against characteristics such as varying image resolutions or attribute distribution seems much more critical for good generalization performance.

**Attributes**. Last, we aim to gather insights about attributes that generalize well and attributes that need further research in order to achieve the performance necessary for practical application in a real-world scenario. For this, we provide mA scores for selected attributes in Fig. 5. In general, gender, hair length, and clothing lengths generalize well to unseen domains. Regardless of the approach used, high mA scores are achieved across all evaluation splits. These attributes are often clearly identifiable and have sufficient training examples on all splits. In addition, we have observed that colors that appear regularly (*e.g.*, black or white) or eye-catching colors (*e.g.*, red or yellow upper-body clothing) were predominantly correctly recognized. As expected, problems occur for strongly imbalanced attributes, *i.e.*, attributes with only a few training samples, attributes with greatly varying attribute distributions, or small and highly-localized attributes such as glasses. So, attributes that are also hard to recognize in a specialization setting. This substantiates the finding from the previous paragraph. Shifted attribute distributions seem to be more important than differences concerning image data shifts. Moreover, the results indicate that the difference between the baseline and *melaeric's* w.r.t. mA arises from the per-

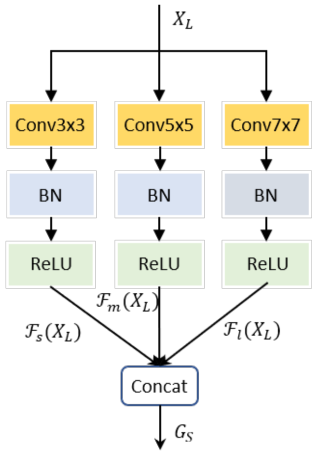

Figure 4: **Spatial Projection from *melaeric*** – Three different sizes of convolution kernels, *i.e.*, 3 × 3, 5 × 5 and 7 × 7 are used to obtain the spatial features of different receptive fields. Each convolutional layer is followed by a Batch Normalization (BN) layer and the ReLU activation. The features from three branches are concatenated together to represent image spatial features.

formance for hard-to-recognize attributes. While the UPAR baseline even outperforms the winning approach regarding well-recognized attributes (gender mA: 89.3% vs. 86.8%), *melaeric* achieves significantly better performance on the complex cases. For instance, the mA for regular glasses is about 7.5% points higher. Similar results are obtained for attributes such as rare lower-body colors (purple, yellow, green, orange) or age, which suffers from imbalanced distributions. The focus on person retrieval of the UPAR baseline leads to the fact that the single attribute recall is low compared to the precision. Since *melaeric* uses higher spatial resolution, spatial projection, and attention, they achieve higher recall scores and therefore mA values.

## 5. Conclusions

The UPAR Challenge attracted over 42 participants, who made 80 submissions during validation and 14 submissions for the test set. Most of the participants could not manage to beat the proposed baseline. Only one participant could surpass the baseline by quite a large margin. Especially for attributes with limited training samples, the participant achieved more robust results than the baseline in terms of resistance to domain gaps. The challenge and these results highlight the difficulty of Pedestrian Attribute Recognition and Attribute-based Person Retrieval in real-world surveillance scenarios. The evaluation protocols for both tasks were challenging and designed to reflect real-world challenges. While 0.6M annotations for the P-DESTRE dataset, with imagery taken from drones, especially contributed to the UPAR dataset for this challenge, the dataset itself did not influence the final results much. Interestingly, while improving the mA metric significantly for the PAR task using a complex combination of Transformer and GCN, the winner solution could not improve against the baseline w.r.t. to the instance-based F1 or in the retrieval task. Analysis of the results reports that although there are more significant domain gaps concerning image data between some datasets, the performance against varying image resolutions and attribute distribution seems much more critical for good generalization performance. Following Specker *et al.* [27] and the results of this challenge, we emphasize that future research should focus more closely on realistic application scenarios instead of smaller individual datasets.

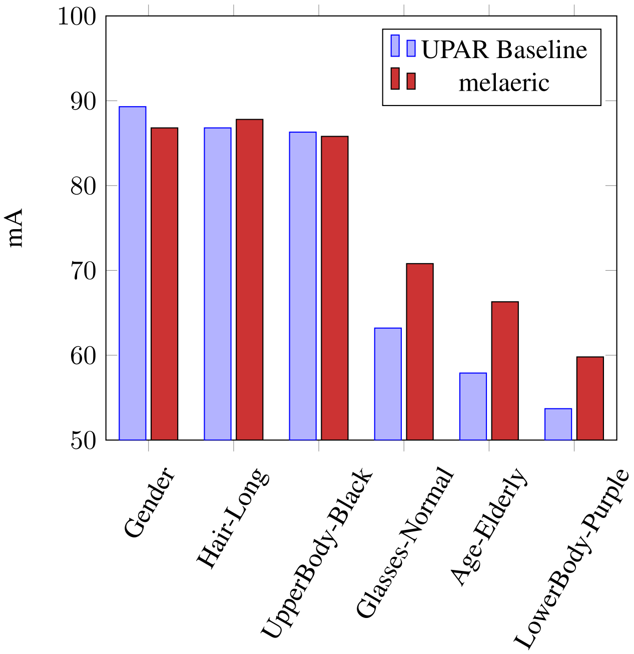

Figure 5: **Cross-domain mA scores** – Comparison of the baseline and the winning approach w.r.t. mA of selected attributes. Values are average scores across all evaluation subsets and splits. The figure indicates that *melaeric*'s approach especially outperforms the baseline concerning hard-to-recognize attributes. Regarding easy attributes, both methdos lead to similar results.

## Acknowledgments

This work has been partially supported by the Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

# References

[1] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. Symbiotic adversarial learning for attribute-based person search. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

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

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

[4] 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, 2021.

[5] 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.

[6] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2020.

[7] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 962–971, 2021.

[8] 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.

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

[10] SV Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, BS Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020.

[11] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115, 2015.

[12] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[13] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2018.

[14] 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, 2019.

[15] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. In *IJCAI*, pages 3177–3183, 2018.

[16] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[17] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018.

[18] 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.

[19] 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.

[20] 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.

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

[22] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022.

[23] 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.

[24] 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.

[25] Arne Schumann, Andreas Specker, and Jürgen Beyerer. Attribute-based person retrieval and search in video sequences. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[26] 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)*, pages 2378–2382. IEEE, 2021.

[27] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[28] 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.

[29] Andreas Specker, Arne Schumann, and Jürgen Beyerer. An evaluation of design choices for pedestrian attribute recognition in video. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2331–2335. IEEE, 2020.

[30] Hao Tan, Zichang Tan, Dunfang Weng, Ajian Liu, Yang Yang, and Jun Wan. Parformer: Vision transformer with relation exploration for pedestrian attribute recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023.

[31] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4997–5006, 2019.

[32] 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.

[33] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *national conference on artificial intelligence*, 2018.

[34] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13055–13064, 2020.

[35] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019.

[36] 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. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[37] Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[39] 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.

[40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.