

# Is Gender “In-the-Wild” Inference Really a Solved Problem?

Tiago Roxo and Hugo Proença, *Senior Member, IEEE*

**Abstract**—Soft biometrics analysis is seen as an important research topic, given its relevance to various applications. However, even though it is frequently seen as a solved task, it can still be very hard to perform in wild conditions, under varying image conditions, uncooperative poses, and occlusions. Considering the *gender* trait as our topic of study, we report an extensive analysis of the feasibility of its inference regarding image (resolution, luminosity, and blurriness) and subject-based features (face and body keypoints confidence). Using three state-of-the-art datasets (PETA, PA-100K, RAP) and five Person Attribute Recognition models, we correlate feature analysis with gender inference accuracy using the Shapley value, enabling us to perceive the importance of each image/subject-based feature. Furthermore, we analyze face-based gender inference and assess the pose effect on it. Our results suggest that: 1) image-based features are more influential for low-quality data; 2) an increase in image quality translates into higher subject-based feature importance; 3) face-based gender inference accuracy correlates with image quality increase; and 4) subjects’ frontal pose promotes an implicit attention towards the face. The reported results are seen as a basis for subsequent developments of inference approaches in uncontrolled outdoor environments, which typically correspond to visual surveillance conditions.

**Index Terms**—Soft Biometrics Analysis, In-the-Wild Gender Inference, Visual Surveillance.

## I. INTRODUCTION

SOFT biometrics analysis has been gaining increased interest over the last years, given its applicability to aid in person identification while requiring little to no cooperation from observed subjects and being robust to low quality data [1], [2]. This interest is also enhanced by the availability of massive amounts of data in wild conditions, provided by surveillance cameras and footage gathered by hand-held devices.

Even though soft biometrics inference is frequently seen as a relatively easy task, this work reports results that point to the opposite. In particular, we focus our extensive analysis on the *gender* trait, posing the following research questions:

- How well solved is *in-the-wild* gender inference?
- What features have the most relevance for gender inference accuracy?
- Given the face importance in gender inference, what conditions justify its use?

Gender information can be retrieved from an image containing the human silhouette and face. As such, the conclusions

yielding from the answers to the above questions might apply to other soft biometrics traits with similar characteristics, such as *age*, *body mass*, or *ethnicity*.

Gender inference *in-the-wild*, such as in surveillance environments, faces various challenges (e.g., occlusions, subjects’ pose, image resolution, and lighting conditions) that strongly decrease the state-of-the-art deep learning-based models’ accuracy. When comparing the conditions between cooperative and *in-the-wild* environments, subjects’ pose and image quality are key discriminating factors that might influence inference accuracy. As such, in this work, we correlate image and subject-based features’ importance with models’ accuracy to perceive the importance of each feature in gender inference. To achieve this goal, we define quantifiable image and subject-based features to categorize the datasets’ quality. We define *resolution*, *luminosity*, and *blurriness* as the most influential image-based features and the *keypoints* (KP) *confidence* of the *face*, *upper*, and *lower body* parts as the subject-based features. The KP data is obtained from Alphapose [3], [4], [5], a state-of-the-art pose estimator.

Given the intent to evaluate gender inference in wild conditions, we use the most challenging Pedestrian Attribute Recognition (PAR) datasets in our experiments. Furthermore, to obtain a generic deep learning model performance evaluation, we gather five state-of-the-art PAR inference models and combine their predictions to obtain a model-independent gender inference. To objectively perceive features’ importance, we correlate them with the models’ combined inference and analyze it using the Shapley value [6], a game-theoretic approach to explain machine learning outputs.

Another important analysis in our work regards the usability of the facial region for gender inference. Given that the face is intuitively one of the most important attributes for gender inference, several models have been proposed using only the face for this task. However, the availability of the facial region in wild situations is not a guaranteed condition, and a subject’s pose might influence its importance for face-based gender inference. To evaluate these concepts, we start by creating challenging (wild) face sets, yielding from frontal images of various PAR datasets; frontal images are obtained from a pose-based image division, using Alphapose. Then, we compare a face-based model’s performance trained in Adience - a face-based gender dataset - and evaluate it on the created face sets. Furthermore, we assess the effect of pose in face-based gender inference accuracy by training the model in *any* or *frontal* pose full-body images and evaluating its performance in face datasets.

According to the above points, the main contributions of

Tiago Roxo is with the University of Beira Interior, Portugal, E-mail: tiago.roxo@ubi.pt.

Hugo Proença is with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Portugal, E-mail: hugomcp@di.ubi.pt.

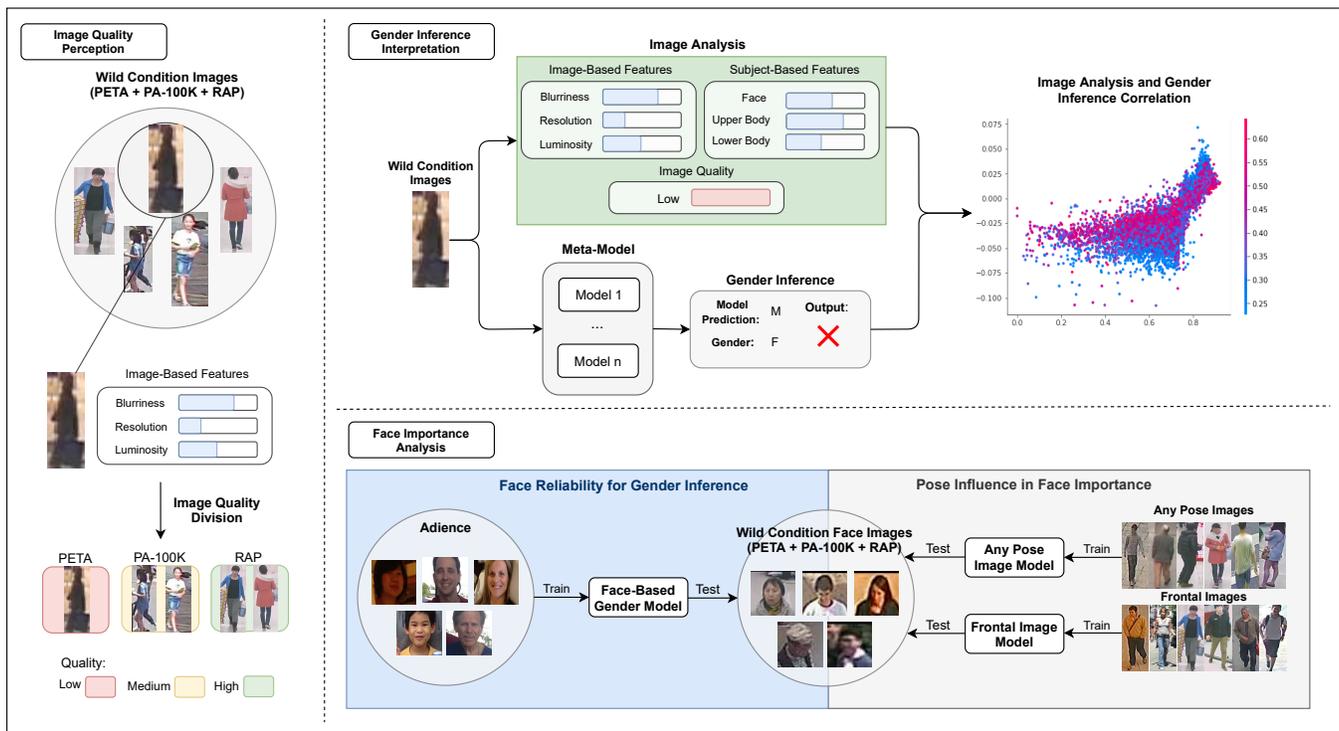


Fig. 1. Overview of the used approaches for image and subject-based features correlation with gender inference. We divide the work into three major phases: 1) image quality perception; 2) gender inference interpretation; and 3) face importance analysis. We divide the PAR datasets considering image-based features and correlate image and subject-based features to gender inference accuracy of a meta-model. This model is a combination of five PAR models. The face importance is analyzed in two ways: by evaluating its reliability in wild conditions and assessing pose influence in face-based gender inference.

this work can be summarized as follows:

- We define the most relevant image and subject-based features for gender inference and relate their importance with image quality;
- We describe an approach to obtain a novel wild face dataset from various PAR sources, using a pose estimation method for head detection and Region of Interest (ROI) definition;
- We evaluate the importance of facial regions for *in-the-wild* gender inference, relating it with image quality;
- We report an implicit face attention mechanism for gender inference, via frontal pose image training.

The remainder of this paper is organized as follows: Section II summarizes the most relevant face and body-based methods used for gender inference; Section III describes the methodologies used for our analysis and Section IV discusses the results obtained. The main conclusions and future work are presented in Section V.

## II. GENDER INFERENCE

### A. Face-Based Approaches

The facial region is commonly used for gender inference, with several datasets and models designed for this task. Levi and Hassner [7] proposed a simple Convolutional Neural Network (CNN) model, with only five layers, given the nature of age and gender inference (small number of classes). Zhan *et al.* [8], used an approach based on Residual Networks of Residual Networks (RoR), which outperformed other CNN

architectures. In the same context, Ranjan *et al.* [9] developed HyperFace, a multi-task framework based on a trunked Alexnet [10] for face detection, fused with a CNN for face landmarks detection and gender recognition.

The facial region usefulness in gender inference has been explored using real-time prediction [11], minimizing the number of parameters by combining *Conv2D* and *BatchNorm* blocks and avoiding *Fully Connected* (FC) layers. Furthermore, Lee *et al.* [12] described a lightweight inference approach that uses depthwise separable convolution layers to reduce the model size and save inference time.

The inference of facial attributes has also been reported in the literature. Liu *et al.* [13] combined a Support Vector Machine (SVM) for face attribute inference with two CNNs: LNET for appropriately resized face localization and ANET for distinguishing face identities, using global and local convolutions for feature extraction. Ryu *et al.* presented InclusiveFaceNet [14], proposing the inclusion of race and gender for face attribute detection, based on modifications of FaceNet [15].

Regarding gender inference based on facial data, several datasets have been announced [13], [16], [17], [18]. All these datasets represent relatively cooperative (controlled) environments, with good image quality, being Adience [19] the most challenging one.

### B. Body-Based Methods

Given the conditions in surveillance scenarios, it is important to consider that the facial region might not be available,

which supports the use of body-based models for gender inference. In this context, PAR is a topic broadly discussed in recent works, aiming to identify pedestrian attributes such as clothing, accessory usage, gender, and age, under challenging covariates such as occlusion, pose, image resolution, and luminosity.

Wang *et al.* [20] proposed a model based on a Recurrent Neural Network (RNN) encoder-decoder framework, using Long Short-Term Memory (LSTM) as a recurrent neuron. They promoted joint recurrent learning by using their framework in images divided into six horizontal strips. Zhao *et al.* [21] also used a similar partition strategy to perform attribute recognition, combining LSTM and BN-Inception [22] networks.

Attention-based approaches are also highly popular: Sarafianos *et al.* [23] combined primary and attention classifiers outputs to learn effective attention maps at multiple scales. Tang *et al.* [24] focused on a feature pyramid structure-based model, with four different feature levels, each with Attribute Localization Modules (ALM). ALM is inspired by Spatial Transformer Network (STN) [25], which discovers the discriminative regions for each attribute in a weakly supervised manner. Guo *et al.* [26] defined a new attention consistency loss, translated by the distance between transformed attention heatmaps of original and flipped images. This approach provides more focused heatmaps, resulting in classification performance improvements.

The analysis of attributes' importance for pedestrian classification is also reported in the literature. Li *et al.* [27] performed global image analysis, exploring how one attribute contributes to the representation of others. They introduced DeepMAR, a learning framework that recognizes multiple attributes simultaneously, exploring the relationships among them. Lin *et al.* [28] developed a re-identification model that considers the attribute importance for this goal. They introduced an Attribute Re-weighting Module (ARM), which corrects attribute predictions based on the learned dependency and correlation between them. Jia *et al.* [29] showed that enhancing localization of attribute-specific areas, typically adopted by state-of-the-art methods, may not be beneficial for performance improvement.

For additional information on this topic, Wang *et al.* [30] review the main divisions of existing PAR approaches, providing a brief summary of the state-of-the-art PAR methods.

### III. METHODOLOGIES

An overview of our processing chain to analyze the effectiveness of gender inference and its correlation with different image features is shown in Fig. 1. We divide three well known PAR datasets (PETA, PA-100K, and RAP) into three image quality levels, based on image-based feature analysis. Then, we correlate image and subject-based features with the gender inference accuracy of a meta-model, in order to perceive the image/subject-based features that favor/penalize overall accuracy. The used meta-model is a combination of the output of five different state-of-the-art PAR models. Furthermore, we analyze the importance of the facial region in gender inference,



Fig. 2. Dataset examples of the evaluated features, divided into image (luminosity and blurriness) and subject-based (head, upper body, and lower body availability) features. Each row corresponds to one of the three PAR datasets. Each column contains opposite cases of each feature, with the left one representing *bad quality* (with feature value closer to its minimum) and the right one representing *good quality*. For the blurriness feature, high feature value represents *bad quality*. All images are resized to the same resolution for visualization purposes.

evaluating a face-based model performance in wild conditions and examining the importance of pose in face-based gender inference.

#### A. Methods and Datasets

For gender inference analysis, we use five models with publicly available implementations. We consider DeepMar [27] and StrongBase [29], which aim to infer pedestrian attributes based on global image analysis, and VAC [26] and ALM [24], which consider attention mechanisms for this task. Additionally, we consider the PAR component of the re-identification method ARP [28]. Regarding the models backbones, all methods are based on ResNet50 [31], except for ALM, which is based on BN-Inception [22].

Regarding the selection of datasets, our goal was to use a set of images with a wide variability range to reproduce, as much as possible, *in-the-wild* conditions. For this reason, we use PETA [32], PA-100K [33], and RAP [34], three state-of-the-art PAR datasets. The chosen datasets clearly display different levels of quality, as discussed in our experiments.

#### B. In-the-Wild Major Variability Factors

Our primary goal was to perceive what is a *good quality* image, in wild conditions, for gender inference. This requires measuring the factors that typically degrade gender inference. We propose a division in image and subject-based features, described in the following subsections. Examples of image and subject-based features analyzed are shown in Fig. 2, displaying contrasting examples of each feature.

1) *Image-Based Features*: We consider *blurriness*, *luminosity*, and *resolution* as the major image-based factors for gender inference. We retrieve image resolution by multiplying its width and height. For luminosity, we obtain the values



Fig. 3. Examples of images considered as *frontal*, *sideways* and *backside*. Each row is related to one of the three PAR datasets and each column regards examples of the respective pose.

of red, green, and blue channels to measure the perceived brightness [35]. Blurriness yields from the convolution of the images with a Laplacian kernel, taking the variance as result.

2) *Subject-Based Features*: We use Alphapose [3], [4], [5], an accurate multi-subject pose estimator, to extract subject-based features. Given that the used datasets contain subjects' ROI, we change the Alphapose implementation to ignore its You Only Look Once (YOLO) [36] detector input. This change translates into a more coherent pose estimation. In exceptional cases, where the image was not trimmed to the subject, we keep the original implementation. The Alphapose outputs 51 KP for each image, corresponding to the image coordinates  $x$  and  $y$ , and the confidence score for each of the 17 Common Objects in Context (COCO) [37] KP: nose, left and right eye, left and right ear, left and right shoulder, left and right elbow, left and right wrist, left and right hip, left and right knee, and left and right ankle. We group the KP into three regions: *face*, *upper*, and *lower body*. *Face* confidence is the mean of nose, eyes, and ears KP confidences. *Upper body* confidence is based on shoulders, elbows, and wrists, while *lower body* confidence uses the values of hips, knees, and ankles.

Regarding subject poses, we consider three values: *frontal*, *sideways*, and *backside*. We differentiate a subject facing forward or backward using pose estimation information. If the rightmost shoulder, from the top left of the image, is the left shoulder, the subject faces forward; otherwise, the subject is facing backward. A subject is considered *sideways* if the shoulder length with respect to the upper body height is less than 0.5. Shoulder length is the absolute difference between the right and left shoulders, and the upper body height is the difference between the shoulder and hip. Some examples of images considered *frontal*, *sideways*, and *backside* are presented in Fig. 3.

### C. Shapley Value

The Shapley value, coined by Shapley in 1953 [6], is a cooperative game theory-based method used for assigning

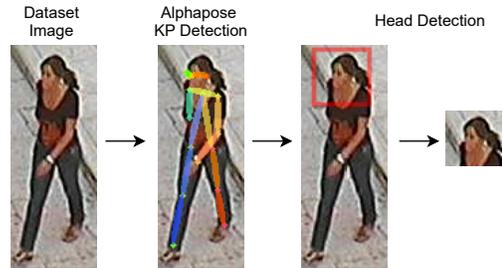


Fig. 4. Phases of the head detection process, displaying pose estimation and head bounding box size. Faces are obtained exclusively from the frontal images of each dataset.

payouts to players, depending on their contribution towards the total payout. In the machine-learning context, the Shapley value is used to evaluate how each feature (player) of a given instance contributed (assigning payout) towards the model prediction of the instance (total payout).

The use of Shapley values in our experiments is linked to our interest in analyzing how each feature contributed to gender inference and if it contributed positively or negatively. The higher the absolute Shapley value is, the higher the feature influence. We associated a value of 1 to instances where all models correctly inferred the gender and 0 otherwise. As such, positive values represent cases that influence correct gender inference, while negative values contribute to gender misinference. We correlate the proposed feature values with gender inference using the SHapley Additive exPlanations (SHAP) framework [38] and the *TreeExplainer* model [39], from a publicly available implementation [40].

### D. Head Detection

Given the varying image quality, we use pose estimation to aid in head detection. Based on the proposed subjects' pose division, we crop the head regions from frontal images. First, we obtain the coordinates  $x$  and  $y$  of the right and left ears and assume that the head's central point is given by the arithmetic mean of such coordinates. Then, head ROIs are drawn using the top-left and bottom-right bounding boxes coordinates, centered in the previously calculated head's central point, always considering the head ROI height as  $\frac{2}{9}$  of the whole body silhouette height. Head bounding boxes drawing is based on Detectron [41]. We provide the steps involved in head detection in Fig. 4, illustrating pose estimation and head ROI definition.

### E. Face Importance Analysis

We evaluate face importance in gender inference by training a face-based gender model in full-body images and obtaining the mean Accuracy ( $mA$ ) [34] in faces of frontal images of the corresponding dataset. The  $mA$  criterion is the mean classification accuracy of the positive and negative samples for each attribute, averaged over all attributes:

$$mA = \frac{1}{2N} \sum_{i=1}^M \left( \frac{TP_i}{P_i} + \frac{TN_i}{N_i} \right), \quad (1)$$

where  $N$  is the number of examples,  $M$  the number of attributes,  $P_i$  is the number of positive examples and  $TP_i$  is the number of correctly predicted positive examples for the  $i^{\text{th}}$  attribute;  $TN_i$  and  $N_i$  are defined analogously.

To obtain an objective measure of face importance ( $FI$ ), we transform the  $mA$  values of frontal image faces ( $mA_f$ ) as follows:

$$FI = \frac{mA_f - 50}{mA_{max} - 50}, \quad (2)$$

where the value 50 represents the model inference randomness, and  $mA_{max}$  is the  $mA$  value obtained when the model is evaluated in full-body images. FI value is enclosed in the  $[0, 100]$  interval, directly corresponding to the face importance in each evaluated context.

## IV. EXPERIMENTS

### A. Dataset and Evaluation Metrics

Our experiments are carried out in three state-of-the-art PAR datasets: PETA [32], PA-100K [33], and RAP [34]. The PETA dataset has 19,000 images, each with 61 binary attributes. It is divided into three subsets: 9,500 for training, 1,900 for validation, and 7,600 for testing. Based on this dataset’s protocol, we report the results on the 35 attributes with a ratio of positive labels higher than 5%. The PA-100K dataset is composed of 100,000 images from outdoor surveillance cameras. It is split into 80,000 images for training, 10,000 for validation, and 10,000 for testing. Each image has 26 attributes. The RAP dataset contains 41,585 images, each annotated with 72 attributes. Following the official protocol [34], we split the dataset into 33,268 training images and 8,317 test images and report the results on the 51 binary attributes with a positive ratio higher than 1%.

For face-based gender inference, we use Adience [19], a dataset composed of face images obtained from cameras, smartphones, and tablets. This dataset contains varying quality factors such as facial expressions, blurred samples, partial occlusions, and head pose variations. It is composed of 26,580 unconstrained images of 2,284 subjects. We use the original in-plane aligned version of the faces [19].

To present our results, we use the label-based metric  $mA$  [34]. For state-of-the-art comparison purposes, we use accuracy when evaluating our face-based gender model (described in Section IV-B2) in the Adience dataset.

### B. Implementation Details

1) *Baseline Methods*: The models used were VAC [26], ARP [28], ALM [24], StrongBase [29], and DeepMar [27].

The VAC model train/test transformation for PETA and RAP is based on the WIDER dataset [42] transformation. Since the attribute weights for weighted sigmoid cross-entropy [27] loss were only defined for PA-100K, we retrieve the attribute weights for PETA and RAP from the same source the authors used to obtain the PA-100K ones. To allow gender inference comparison in our experiments, we change the PETA dataset

evaluation for the StrongBase evaluation implementation, adjusting the output-related loss without modifying attention-related ones.

For ARP, we only consider attribute loss, ignoring the re-identification one. As the available implementation did not consider PETA, RAP, and PA-100K dataset evaluation, we introduce it following the original paper’s guidelines.

The remaining models are evaluated using the original implementations. All the gender models used in our experiments are specifically trained for gender attribute only. For all models’ implementations that did not consider the evaluated datasets, we incorporate StrongBase dataset reading into their implementation. For consistency purposes, all metric evaluations are also based on StrongBase’s implementation.

2) *Face-Based Gender Model*: The proposed model is based on the StrongBase implementation [29]. The ResNet50 [31], pretrained in ImageNet [43], is used as backbone to extract image features. The model is separated into feature extraction and classification. The feature extractor is composed of ResNet50 without its *Average Pool* and *FC* layer, which were incorporated in the classifier. The classifier has its *FC* layer replaced by a *Linear* layer, followed by a *Batch Normalization* one. Images are resized to  $256 \times 192$  with random horizontal mirroring as inputs. Stochastic Gradient Descent (SGD) is used for training, with momentum of 0.9, and weight decay of 0.0005. The initial learning rate is set to 0.01 for the feature extractor and 0.1 for the classifier. Plateau learning rate scheduler is used with a reduction factor of 0.1 and a patience epoch number of 4. Batch size is set to 64 and the total epoch number of training is 30, using Binary Cross Entropy with Logits (BCELogits) as loss function.

### C. Cross-Domain Performance

We evaluate the gender inference performance of each model, when trained for all pedestrian attributes, in the *within* and *cross-domain* settings. This analysis allows us to perceive the potential robustness of each model. Additionally, we train each model specifically for the gender attribute and compare its performance with the corresponding model trained for all dataset attributes. We report our results in Table I.

From this experiment, we can observe that models trained exclusively with the gender label tend to have slightly better gender inference accuracy than those trained for all attributes. This is confirmed in both *within* and *cross-domain* scenarios. Therefore, in all the remaining experiments, we resort to models trained exclusively for gender.

Regarding the obtained models’ accuracies, we conclude that no model is significantly better than the others, regardless of the dataset where it was trained/evaluated. However, when examining model performance within and across datasets, some differences can be observed. First, models trained in PETA or PA-100K and evaluated in other datasets tend to perform well, achieving results close to those they present when evaluated in the same dataset they were trained on. The exception is RAP-trained models, where they perform considerably worse if evaluated in other datasets. Second, considering the *within-domain* setting, RAP is the dataset

TABLE I

GENDER  $mA$  OF DIFFERENT MODELS TRAINED AND EVALUATED ON VARIOUS PAR DATASETS. MODELS NAMED WITH *Gender* SUFFIX REFER TO MODELS TRAINED SOLELY FOR THE GENDER ATTRIBUTE. THE OUTPERFORMING METHODS FOR EACH DATASET ARE SHOWN IN BOLD.

Train \ Eval	Methods	PETA	PA-100K	RAP
PETA	StrongBase	92.80	78.98	79.47
	StrongBase-Gender	<b>93.13</b>	79.65	80.44
	ALM	91.36	76.95	80.62
	ALM-Gender	92.28	78.27	82.66
	DeepMar	91.08	72.83	71.76
	DeepMar-Gender	92.33	77.63	74.78
	VAC	92.85	78.16	78.50
	VAC-Gender	92.76	78.22	77.92
PA-100K	APR	92.84	78.83	78.95
	APR-Gender	92.34	75.49	78.49
	StrongBase	77.28	90.97	86.93
	StrongBase-Gender	75.01	90.77	88.38
	ALM	75.48	88.34	85.87
	ALM-Gender	76.59	90.34	88.55
	DeepMar	76.17	88.52	85.73
	DeepMar-Gender	74.50	90.39	85.72
RAP	VAC	75.20	90.17	87.79
	VAC-Gender	76.85	<b>91.05</b>	88.98
	APR	76.53	89.91	87.24
	APR-Gender	75.26	90.05	86.29
	StrongBase	67.67	72.25	<b>96.74</b>
	StrongBase-Gender	71.41	76.97	96.34
	ALM	76.81	81.24	95.69
	ALM-Gender	76.03	80.28	95.41
PETA	DeepMar	69.29	75.76	95.52
	DeepMar-Gender	70.22	76.89	96.40
	VAC	66.11	68.84	96.52
	VAC-Gender	72.89	76.67	96.59
	APR	68.39	75.41	96.20
	APR-Gender	69.24	72.93	95.90

where models perform the best, achieving the highest accuracy out of all the three evaluated datasets. Finally, models trained in PETA or PA-100K achieve better performances in RAP than in any other dataset, aside from the one they were trained on. PETA is the dataset where models, if trained on other datasets, perform the worst. These observations indicate that accurate gender inference is easier in RAP images and that PETA is the most challenging dataset for this task.

#### D. Dataset Analysis

To perceive how image-based features influence gender inference, we obtain the corresponding values for each feature/dataset. We present the normalized mean and standard deviation values in Table II.

Given the low variability of luminosity across the datasets, we group them into categories exclusively based on resolution and blurriness: *low*, *medium*, and *high*. The higher quality dataset is RAP, given the high resolution value and sharpness. PETA is the dataset with lower quality, based on the low resolution values and high blurriness. Since PA-100K dataset feature values are typically between those of PETA and RAP, we consider this dataset as of medium quality. We present a representative quality bar of the three datasets in Fig. 5, illustrating some dataset examples.

TABLE II

DATASET IMAGE-BASED FEATURE ANALYSIS. VALUES ARE NORMALIZED FOR THE COMBINATION OF THE THREE DATASET VALUES, FOR EACH FEATURE.

Dataset	Resolution	Luminosity	Blurriness
PETA	$0.037 \pm 0.024$	$0.432 \pm 0.100$	$0.120 \pm 0.148$
PA-100K	$0.062 \pm 0.063$	$0.449 \pm 0.126$	$0.095 \pm 0.089$
RAP	$0.131 \pm 0.083$	$0.407 \pm 0.107$	$0.022 \pm 0.013$

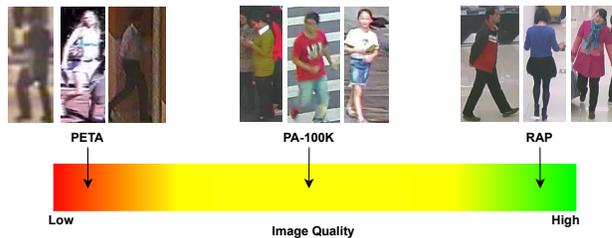


Fig. 5. PAR dataset quality division, taking into account image-based feature values. Each dataset illustrates examples of the corresponding quality levels.

#### E. Methods Evaluation

We evaluate the image and subject-based features of the correctly inferred cases, in comparison to all test set images, to perceive the features that contribute the most to gender misinference. Here, we consider the *within-domain* setting. Given the predictions of the five methods used, we consider *correctly classified* cases exclusively when all models correctly infer the corresponding value. Results are displayed in Table III, with values normalized per dataset.

The results suggest no significant differences between images where gender is correctly inferred and all test set images, for all the evaluated features. This experiment points out that a general feature evaluation of the correctly gender-inferred images might not be enough to identify how each feature contributes to gender inference accuracy.

#### F. Interpreting Gender Inference

With the intent to interpret feature values of different quality images and their relation to gender inference, we grouped all models' predictions, for all datasets, and analyze their correlation with gender inference, as described in Section III-C.

By analyzing the bar plot for each feature mean absolute SHAP values, in Fig. 6, we conclude that image resolution and lower body KP confidence are the most important characteristics. Bars in blue are positively correlated values, while the red bar denotes a negatively correlated feature. In this case, high values of blurriness contribute to gender misinference. The resolution importance was expected *a priori*, given the difficulties in discriminating gender in poor resolution data. The lower body importance might be related to hip or footwear influence, which is likely to occur when the face is fully occluded, forcing models to analyze different parts for gender inference. Subjects' pose is the least influential factor among those evaluated.

To better understand each feature's importance in gender inference, we evaluate the dependence plots for some features.

TABLE III

IMAGE AND SUBJECT-BASED FEATURE VALUES FOR PETA, PA-100K, AND RAP DATASETS. *Correct* REPRESENTS THE CORRECTLY INFERRED IMAGES (WHEN ALL MODELS INFERRED THE CORRECT GENDER VALUE) AND *All* REFERS TO ALL TEST DATA. SUBJECTS POSE VALUES REPRESENT THE PORTION OF TEST IMAGES WITH THE GIVEN POSE.

Features		PETA		PA-100K		RAP	
		Correct	All	Correct	All	Correct	All
Subject Pose	Frontal	0.387	0.383	0.363	0.355	0.350	0.348
	Sideways	0.223	0.250	0.356	0.368	0.309	0.315
	Backside	0.390	0.366	0.282	0.277	0.341	0.337
KP Confidence	Face	0.823 ± 0.105	0.821 ± 0.107	0.820 ± 0.124	0.813 ± 0.128	0.839 ± 0.117	0.835 ± 0.125
	Upper Body	0.781 ± 0.109	0.775 ± 0.114	0.773 ± 0.099	0.765 ± 0.104	0.820 ± 0.094	0.817 ± 0.098
	Lower Body	0.764 ± 0.108	0.759 ± 0.112	0.768 ± 0.087	0.761 ± 0.093	0.786 ± 0.115	0.783 ± 0.118
Image	Resolution	0.211 ± 0.131	0.207 ± 0.132	0.066 ± 0.120	0.054 ± 0.070	0.223 ± 0.157	0.227 ± 0.155
	Luminosity	0.417 ± 0.116	0.415 ± 0.119	0.452 ± 0.160	0.462 ± 0.158	0.478 ± 0.152	0.475 ± 0.153
	Blurriness	0.148 ± 0.188	0.143 ± 0.175	0.114 ± 0.120	0.110 ± 0.118	0.115 ± 0.070	0.114 ± 0.070

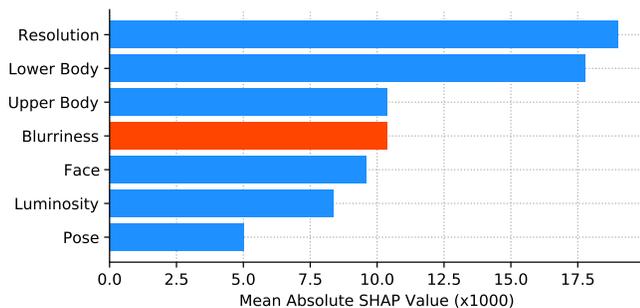


Fig. 6. Bar plot of the mean absolute SHAP values for image and subject-based features. Blue colored bars correlate with gender correct inference, while the red one correlates with gender misinference. Features are ordered by its influence in gender inference accuracy.

Fig. 7 shows the results for the face, lower, and upper body KP confidence (subjects-based factors) and resolution, luminosity, and blurriness dependence plots (image-based features). Negative SHAP values are linked to gender misinference, while positive ones are linked to correct inference. The face and lower body KP confidence plots correlate each corresponding feature with lower body and luminosity, respectively. For the plots with no feature correlation, the color variation from blue to red represents an increase in feature values. We conclude that higher KP confidence values translate into gender correct inference. For lower and upper body, KP confidence values above 0.7 are linked to correct inference, while face requires KP confidence values above 0.8 for the same effect. Low values of resolution heavily negatively impact gender inference and are positively influential in higher values. Luminosity only influences gender inference when it has low values and tends to have a neutral role with value increase. Blurriness has a positive influence on low values and a negative influence when its value arises. An increase in blurriness does not translate into more incorrect gender inference.

Regarding the feature correlation plots, for face confidence values lower than 0.8, higher confidence of lower body tends to translate into higher SHAP values. This indicates that models tend to focus on the lower body portion to perform inference when the face is not visible. Regarding the lower body, increased luminosity generally improves the accuracy, which is more evident when KP confidence is between 0.4

and 0.7. This suggests that when a subject’s lower body is in uncooperative poses or minor occluded, better illumination might dissipate the difficulty in interpreting such poses for gender models.

The observed results suggest that higher resolution images with good illumination, low blurriness, and cooperative poses (translated in high KP confidence values) are important factors for correct gender inference. Furthermore, the face might not be reliable for *in-the-wild* gender inference since it requires higher confidence values than those of the lower and upper body for accurate inference.

In addition to the overall analysis of features influence, we further investigate if image quality can change the importance of each image/subject-based feature. For this, we analyze feature importance in each dataset, translated in low (PETA), medium (PA-100K), and high (RAP) quality images. We present the bar plots of our analysis in Fig. 8. All SHAP values refer to the mean absolute values of each feature.

We can conclude that image-based features are highly influential in low quality images, given the high SHAP values they present in these conditions. This analysis also suggests that subject-based features are not reliable for gender inference in scenarios where image quality is low. If we examine image-based features across image quality variance, we conclude that they tend to have less importance as quality increases. Subject-based features increase their importance from low to medium quality images. Additionally, in this quality range, subject-based features are more important than image-based, illustrating a shift of feature importance for gender inference. In high quality images, the SHAP values are all lower than in other quality range images. This indicates that each feature loses importance in high quality data, where the models can analyze different features reliably. The pose is the least influential feature, presenting an insignificant influence in high quality images.

These findings can be applied to improve model inference. Our results suggest that, if the inference is done using low quality images, it is advised to collect new data with improved image-based features. However, assuming medium quality data usage, image quality increase (such as having better illumination or resolution) is not as important as improving subject-based features. In high quality images, gender inference accuracy improvements might not be achieved through

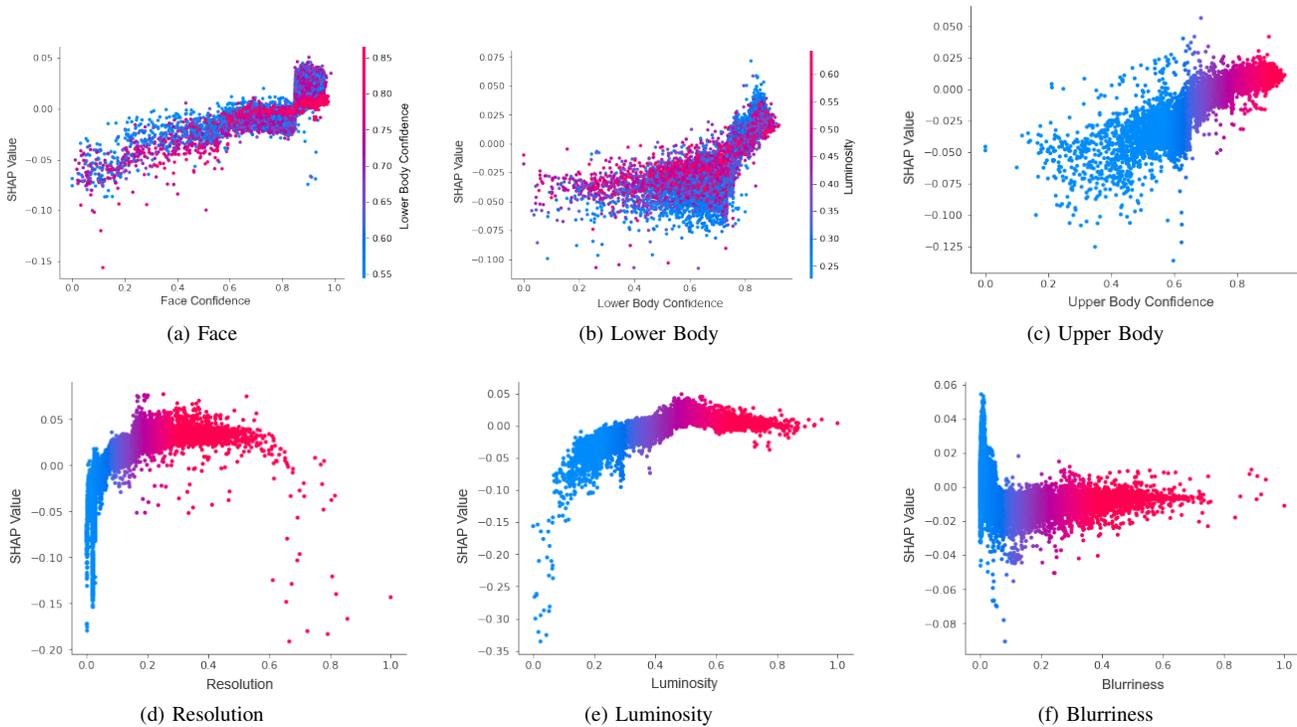


Fig. 7. Dependence scatter plots of Face (a), Lower Body (b), Upper Body (c), Resolution (d), Luminosity (e), and Blurriness (f) SHAP values. Face and Lower Body plots present a correlation with the named feature and Lower Body and Luminosity, respectively. These plots contain a right  $y$  axis with color variation, representing correlated feature value variance. Color variation from blue to red in the other plots represent increasing values of the feature. Image-based features are normalized.

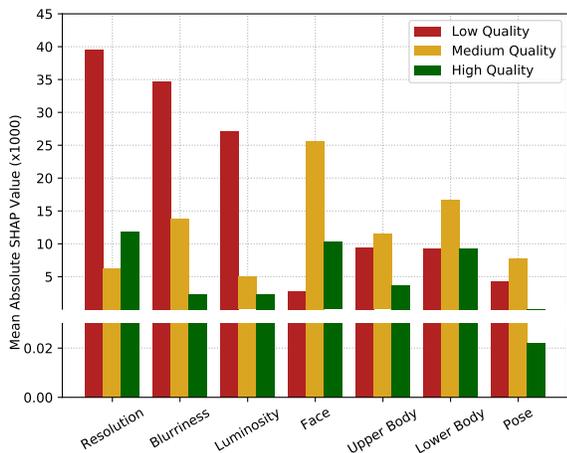


Fig. 8. Bar plots of the evaluated features, presenting the SHAP values for low, medium, and high quality image datasets. Datasets are represented by different colors. High SHAP values represent high feature relevance for gender correct inference. The bar plot is broken in the vertical axis for better visualization.

data changes, given the relatively low influence of the analyzed features for gender inference. Redefining models' architecture might be a better strategy to improve overall accuracy.

### G. Face Importance in Gender Inference

The results described in section IV-F suggest that face importance for gender inference might be directly related to

TABLE IV  
COMPARISON OF THE PROPOSED FACE-BASED GENDER MODEL, IN THE ADIENCE DATASET.

Methods	Acc (%)
Lee <i>et al.</i> (2018) [12]	85.16
Levi and Hassner (2015) [7]	86.80
Hung <i>et al.</i> (2019) [44]	89.50
Zhang <i>et al.</i> (2017) [8]	92.43
Face-Model	93.79

image quality. To confirm this idea, we analyze the reliability of face-based gender inference given different data quality. We start by training our face-based gender model in Adience, one of the most challenging face gender datasets. We present the accuracy obtained for our model in Table IV, in comparison to results reported in the literature (5-fold cross-validation [19] values).

The obtained accuracy is better than the reported ones, indicating that our model represents a good face-based model. To assess its applicability in wild conditions, we evaluate its *cross-domain* performance using Adience, PETA, PA-100K, and RAP datasets. To keep the Adience values' comparison as fair as possible, we train and evaluate the models in PETA, PA-100K, and RAP using only face images of frontal subjects of each dataset. We report the results in Table V. The Adience result is based on the proposed face-based gender model when training and testing on the first folder of the 5-fold cross-validation.

We can observe that the Adience-trained model performs

TABLE V

GENDER  $mA$  OF FACE-BASED GENDER MODEL, TRAINED AND EVALUATED IN WITHIN AND CROSS-DOMAIN SETTINGS. ONLY FRONTAL FACE IMAGES OF PETA, PA-100K, AND RAP DATASETS WERE CONSIDERED.

<b>Train \ Eval</b>	Adience	PETA	PA-100K	RAP
Adience	94.83	65.56	55.55	77.85
PETA	64.96	91.16	74.60	82.43
PA-100K	75.62	72.52	84.71	79.72
RAP	75.49	68.09	68.32	95.31

TABLE VI

GENDER  $mA$  OF FACE-BASED GENDER MODEL, TRAINED FOR *Any* AND *Frontal* POSE IMAGES, AND EVALUATED IN FRONTAL IMAGE FACES (*Face*) AND FULL-BODY IMAGES OF CORRESPONDING POSE (*Body*). *Any Pose* REFERS TO ALL DATASET IMAGES AND *Frontal Pose* IMAGES REFER TO FRONTAL IMAGES. FI STANDS FOR FACE IMPORTANCE.

<b>Dataset</b>	<b>Any Pose</b>			<b>Frontal Pose</b>		
	Face	Body	FI (%)	Face	Body	FI (%)
PETA	57.60	93.24	17.58	61.09	92.52	26.08
PA-100K	52.52	91.06	6.14	57.03	91.86	16.79
RAP	73.63	96.09	51.27	75.31	95.98	55.05

poorly in other datasets, illustrating that face-based training is not adequate for *in-the-wild* inference. However, the dataset where this model performs the best is RAP, characterized as a high quality images set. Similarly, RAP-trained models achieve higher accuracy performance in Adience than in PETA or PA-100K. These results indicate that Adience and RAP face dataset have similarities, which corroborates that face importance is closely related to increased image quality. To further explore this concept, we assess image quality and the subject’s pose effect on face importance. For this, we train our face-based gender model in all dataset images (full-body) and evaluate its performance in face images of the correspondent dataset. The procedure is repeated when training the model only in frontal images. We present the obtained results in Table VI.

The results observed support that higher quality images directly influence the importance of face in gender inference. Furthermore, when we train the model with frontal data and evaluate using only facial images, the face importance increased for all three datasets, suggesting that the face importance factor is also linked to its availability. This finding suggests an implicit attention towards the face, motivated solely by guaranteed face access, via training in frontal images only.

## V. CONCLUSIONS

Soft biometrics inference in wild conditions is hindered by factors such as partial occlusions, subjects’ pose, and varying levels of image quality. Considering, in particular, the gender trait, this work analyzes the image and subject-based features that are the most/least linked to correct inference of state-of-the-art models. In our analysis, we assess the combined gender inference of five state-of-the-art PAR models in well known PAR datasets. We evaluate the effectiveness of face-based gender inference models in surveillance scenarios, also considering the effect of pose. Our results suggest that image-

based features are more important in low-quality images, while the importance of subject-based features is linked to an increase in image quality. Moreover, the relevance of the facial region to correctly infer gender is directly related to high-quality data. Finally, we show an implicit face attention approach, linked to its guaranteed availability, via frontal image training.

Our findings provide the basis for designing more robust gender inference models, particularly suited to work *in-the-wild*, where image and subject-based features highly vary. The improvement of face-based gender inference via an implicit face attention mechanism paves the way to incorporate similar approaches in future works.

## ACKNOWLEDGMENT

This work is funded by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the project UIDB//50008/2020. Also, it was supported by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT, and supported by FCT - Fundação para a Ciência e Tecnologia through the research grant 2020.09847.BD.

## REFERENCES

- [1] A. K. Jain, S. C. Dass, and K. Nandakumar, “Soft biometric traits for personal recognition systems,” in *International conference on biometric authentication*. Springer, 2004, pp. 731–738.
- [2] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? a survey on soft biometrics,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2015.
- [3] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” in *ICCV*, 2017.
- [4] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, “Crowdpose: Efficient crowded scenes pose estimation and a new benchmark,” *arXiv preprint arXiv:1812.00324*, 2018.
- [5] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, “Pose Flow: Efficient online pose tracking,” in *BMVC*, 2018.
- [6] L. S. Shapley, “A value for n-person games,” *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.
- [7] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.
- [8] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, “Age group and gender estimation in the wild with deep ror architecture,” *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017.
- [9] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [11] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, “Real-time convolutional neural networks for emotion and gender classification,” *arXiv preprint arXiv:1710.07557*, 2017.
- [12] J.-H. Lee, Y.-M. Chan, T.-Y. Chen, and C.-S. Chen, “Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 162–165.
- [13] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

- [14] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusiveness: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017.
- [15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [16] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018.
- [17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.
- [18] S. Escalera, M. Torres Torres, B. Martínez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–8.
- [19] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.
- [20] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 531–540.
- [21] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning," in *IJCAI*, 2018, pp. 3177–3183.
- [22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [23] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 680–697.
- [24] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4997–5006.
- [25] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.
- [26] H. Guo, K. Zheng, X. Fan, H. Yu, and S. 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*, 2019, pp. 729–739.
- [27] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 111–115.
- [28] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.
- [29] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," *arXiv preprint arXiv:2005.11909*, 2020.
- [30] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," *arXiv preprint arXiv:1901.07474*, 2019.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [32] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792.
- [33] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 350–359.
- [34] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.
- [35] A. Morgand and M. Tamaazousti, "Generic and real-time detection of specular reflections in images," *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, vol. 1, pp. 274–282, 01 2014.
- [36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- [39] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [40] S. e. a. Lundberg, "Shap (shapley additive explanations)," 2021, [Online] <https://github.com/slundberg/shap>. Last access: 14/03/2021.
- [41] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2021, [Online] <https://github.com/facebookresearch/detectron2>. Last access: 16/03/2021.
- [42] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700.
- [43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [44] S. C. Hung, J.-H. Lee, T. S. Wan, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 339–343.