YinYang-Net: Complementing Face and Body Information for Wild Gender Recognition

TIAGO ROXO, (Member, IEEE), AND HUGO PROENÇA, (Senior Member, IEEE)

Instituto de Telecomunicações (IT), University of Beira Interior, 6201-001 Covilhã, Portugal

Corresponding author: Tiago Roxo (tiago.roxo@ubi.pt)

ABSTRACT Soft biometrics inference in surveillance scenarios is a topic of interest for various applications, particularly in security-related areas. However, soft biometric analysis is not extensively reported in *wild conditions*. In particular, previous works on gender recognition report their results in face datasets, with relatively good image quality and frontal poses. Given the uncertainty of the availability of the facial region in wild conditions, we consider that these methods are not adequate for surveillance settings. To overcome these limitations, we: 1) present frontal and *wild* face versions of three well-known surveillance datasets; and 2) propose YinYang-Net (YY-Net), a model that effectively and dynamically complements facial and body information, which makes it suitable for gender recognition in wild conditions. The frontal and *wild* face datasets derive from widely used Pedestrian Attribute Recognition (PAR) sets (PETA, PA-100K, and RAP), using a pose-based approach to filter the frontal samples and facial regions. This approach retrieves the facial region of images with varying image/subject conditions, where the state-of-the-art face detectors often fail. YY-Net combines facial and body information through a learnable fusion matrix and a channel-attention sub-network, focusing on the most influential body parts according to the specific image/subject features. We compare it with five PAR methods, consistently obtaining state-of-the-art results on gender recognition, and reducing the prediction errors by up to 24% in frontal samples. The announced PAR datasets versions and YY-Net serve as the basis for *wild* soft biometrics classification and are available in here.

INDEX TERMS Gender recognition, selective attention, soft biometrics analysis, visual surveillance.

I. INTRODUCTION

Soft biometric cues are particularly useful in security-related issues and have increased in popularity in recent years. This is mainly motivated by their applicability in person identification, while requiring no cooperation from the subjects and being robust to low quality data [1], [2]. Furthermore, the availability of massive amounts of data acquired in uncontrolled conditions, provided by surveillance cameras and hand-held devices, has also contributed to enhancing the interest in this topic.

Gender is a particularly relevant soft label and its classification has been extensively reported in the literature, with various datasets announced [3]–[7]. However, we consider that gender recognition in *wild conditions* is not yet extensively explored, with most of the previous methods frequently

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

linking it to face analysis in relatively good quality data, without the challenges found in surveillance scenarios. Alternatively, gender information can be retrieved in video surveillance settings, using Pedestrian Attribute Recognition (PAR) datasets, with methods analyzing up to 51 attributes simultaneously and not the gender alone. This promotes the use of weighted losses for the different attributes and is an evident obstacle towards the optimization of results for each one.

Given that the face is intuitively a discriminative attribute to classify gender, our main goal is to develop an integrated framework that uses facial information, whenever possible, in surveillance conditions. This is useful for scenarios where analysis is carried out mostly from frontal poses, such as airports, casinos, and border services. However, reliably retrieving the facial region from surveillance images can not be considered a solved problem. In this context, we start by creating *frontal* versions of PAR datasets (PETA, PA-100K, and RAP) using pose data from Alphapose [8]–[10].
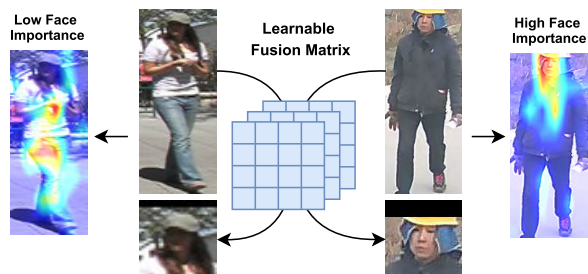
FIGURE 1. Illustration of the effect of the learnable fusion matrix, a major component of YY-Net, as an attention mechanism with respect to the (non)availability of facial information.

Then, we create *wild* face datasets using a pose-based approach to define the head Region of Interest (ROI), from frontal PAR images. This approach is applicable in contexts with varying image quality and uncooperative subjects, outperforming state-of-the-art face detectors. Based on the varying image/subject qualities and face variability factors, the created *wild* face datasets contrast with the state-of-the-art, showing its applicability for wild soft biometrics classification.

To effectively classify gender, using the created datasets, we propose YinYang-Net (YY-Net), a model that combines face and body-based features through a learnable fusion matrix, promoting weighted conjunction of both types of information. Fig. 1 displays the effect of that matrix, with respect to the availability of face information. Moreover, we apply a residual link with body features, using a channel attention sub-network, to promote a weighted face-body combination. This contributes to better focus on the most influential body parts for gender recognition.

According to our experiments, the proposed model improves gender recognition results in all frontal PAR datasets, when compared to five state-of-the-art PAR methods. Furthermore, in non-frontal images, where the facial region is unavailable, YY-Net uses body information and not the weighted combination of face-body features. In this context, it also outperforms the baseline PAR methods, in all datasets.

In summary, the major contributions of this paper are:

- We present frontal and *wild* face versions of well-known PAR datasets (PETA, PA-100K, and RAP), released as meta-data, yielding from a pose processing approach. Given their face variability factors, these *wild* face datasets contrast with the state-of-the-art, representing more challenging sets to soft biometrics applications;
- We propose YY-Net, an attention-based model for *wild* gender recognition, complementing face and body information, that surpasses the state-of-the-art in PAR datasets, particularly in frontal samples.

The remainder of this paper is organized as follows: Section II summarizes the most relevant gender recognition works; Section III describes the proposed model; Section IV describes the proposed datasets, and Section V discusses the

results obtained. The main conclusions and future work are presented in Section VI.

II. RELATED WORK
A. FACE-BASED GENDER RECOGNITION

The face is one of the most discriminative aspects to classify gender, which culminated in several datasets designed for this task [3]–[6]. Given the nature of gender and age estimation, simple Convolutional Neural Network (CNN) models (with only five layers) can be used for gender recognition [11]. However, more complex architectures have recently been used, such as those based on Residual Networks of Residual Networks (RoR) [12], which have outperformed other CNN architectures.

Some works have focused on complementing the face images with facial attributes analysis. Han *et al.* [13] proposed a Deep Multi-Task Learning (DMTL) framework for joint estimation of face attributes, considering their correlation and heterogeneity for various demographic attributes (gender, age, race). Ranjan *et al.* [14] developed HyperFace, a multi-task framework based on Alexnet [15] for face detection, fused with a CNN for face landmarks detection and gender recognition. Liu *et al.* [3] combined a Support Vector Machine (SVM), for face attribute classification, with a CNN for appropriately resized face localization and another for distinguishing face identities.

The complement of particular face attributes for gender classification is also a topic of research. Datcheva and Brémond [16] analyzed smile-face dynamics in gender estimation, relating its importance with the subject's age and demonstrating the complement of dynamic and appearance-based features. Lapuschkin *et al.* [17] examined which facial attributes actually impact age and gender prediction, using a Layer-wise Relevance Propagation. In the same topic, Yaman *et al.* [18] used ear appearance to aid gender estimation in profile face images, employing domain adaptation. Ryu *et al.* presented InclusiveFaceNet [19], proposing the inclusion of race and gender for face attribute detection.

B. BODY-BASED GENDER RECOGNITION

Regarding wild gender recognition, where the focus is on the whole body and not on the facial region alone, we present works using PAR datasets [20]–[22], a context that more closely relates to surveillance scenarios. Some works opt to analyze the relation of different body parts: Wang *et al.* [23] proposed a model based on a Recurrent Neural Network (RNN) encoder-decoder framework, using Long Short-Term Memory (LSTM) as a recurrent neuron, with joint recurrent learning via images division into six horizontal strips; Zhao *et al.* [24] also used a similar partition strategy, combining LSTM and BN-Inception [25] networks.

The importance of various attributes is also reported in the literature, namely on the relation of attributes towards pedestrian attribute recognition. Learning the relationship, dependency, and correlation between attributes and using this
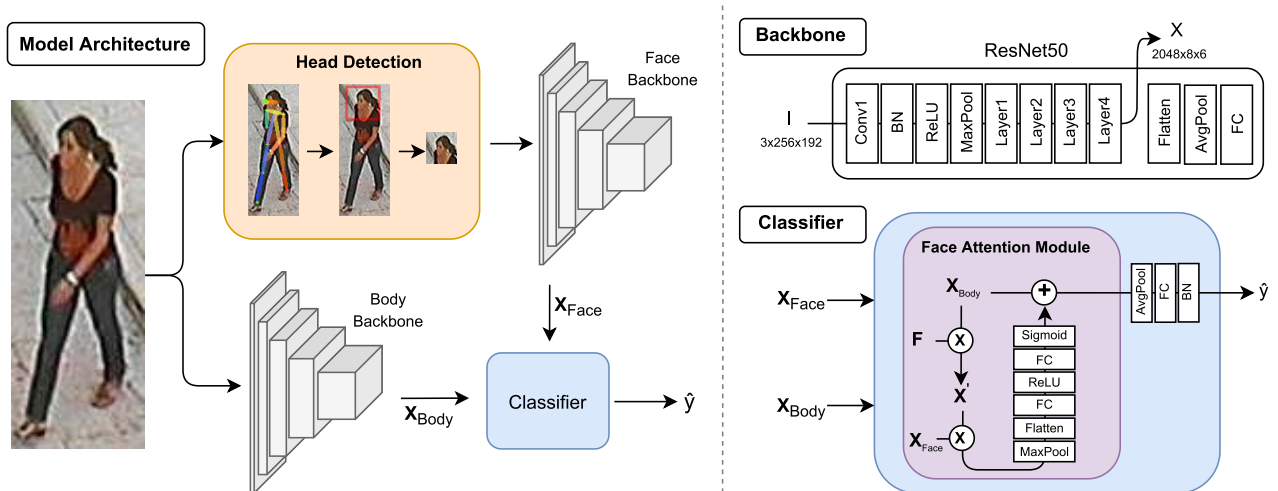
FIGURE 2. Cohesive perspective of the YY-Net, divided into three parts: *1) Model Architecture; 2) Backbone;* and *3) Classifier*. The *Model Architecture* displays our processing flow, where the input data feeds two different backbones: 1) for the body; and 2) for facial information. The facial regions are obtained according our head detection approach, using pose information. The backbones derive from ResNet50, with classifier-related layers ignored, as shown in the *Backbone* portion. The combination of the face and body inputs is done through a classifier, which uses the proposed Face Attention Module to fuse both types of information effectively.

information to accurately predict attributes have been the focus of some works [26], [27]. Similarly, using a hierarchical feature embedding framework [28] has proven to promote finer-grained clustering, translating into increased attribute recognition accuracy. However, enhancing the localization of attribute-specific areas, typically adopted by state-of-the-art methods, may not necessarily improve performance [29]. Image and subject-based features are also aspects to consider for accuracy improvement, given that their importance in gender classification varies with image quality and face availability [30].

Pose information and body part localization are also influential for pedestrian attribute recognition. Li *et al.* [31] used pose data to aid in attribute body part localization, fusing features at multiple levels for attribute recognition. Zhang *et al.* [32] proposed a Deep Template Matching method to capture body parts features, complemented by pose keypoints to guide discriminative cues learning. Liu *et al.* [33] used body ROI and assigned attribute-specific weights based on extracted ROI proposals and attribute localization. Yaghoubi *et al.* [34] used body keypoint estimation to define attribute region of interest, while also evaluating pose effect. Finally, Barra *et al.* [35] explored the usage of body pose information for gender classification in setups with subjects and camera in movement.

C. ATTENTION-BASED GENDER RECOGNITION
Various works base their approaches on attention mechanisms. Effective attention maps at different scales [36], analyzing relevant image patches [37], or focusing on different feature levels [38] are all reported methodologies to discover discriminative regions for attribute classification. Alternatively, using an attention consistency loss [39] based on

heatmaps difference of original and flipped images can yield good performances.

Attribute relation was also the basis of attention approaches [40]–[44]. Li *et al.* [40] used Convolutional LSTM (ConvLSTM) for spatial and semantic correlation between attributes, combining it with channel attention to adjust the weight of relevant channel features adaptively. Zhao *et al.* [42] proposed the use of a Recurrent Attention Model (also based on ConvLSTM), mining the correlations among different spatially related attributes. Ji *et al.* [43] announced an encoder-decoder with multiple LSTM, aiming to exploit more contextual knowledge, mining deeper relations between images and attributes. Gao *et al.* [44] proposed a two-stage training, complementing semantic segmentation with attribute recognition, relating attributes in an attention-based manner between both stages.

III. YinYang-Net
An overview of YY-Net is shown in Fig. 2. The input is processed in two ways: 1) retrieving body information; and 2) retrieving face information. We obtain facial information by first processing the image to obtain pose data and then using a head ROI detection approach, described in Section IV-A. We use Alphapose [8]–[10] to collect pose information. Both body and face information are retrieved from a modified ResNet50 [45], via feature maps at *Layer 4*. Finally, we pass both feature maps to a classifier for gender recognition.

A. NETWORK ARCHITECTURE
The key concept of our approach is to combine face and body information to classify gender. The facial region is intuitively a discriminative attribute for gender recognition, and its importance is linked to the subject pose. To achieve

our goal, we use a ResNet50, pretrained in ImageNet [46], for feature extraction of both face and body images. To better control features combination, we retrieve features from a modified ResNet50 model, with *Flatten*, *Pooling*, and *FC* layers removed. We refer the modified ResNet50 as *backbone*. We use two different backbones for face and body image processing to fine-tune extraction based on the body region inputted.

Given an input image I, we denote the features extracted as $\phi_i(I) \in \mathbb{R}^{H_i \times W_i \times C_i}$, $i \in \{\text{face}, \text{body}\}$. For a $3 \times 256 \times 192$ input image, we obtain a $H_i \times W_i \times C_i$ of $2048 \times 8 \times 6$, for both values of i. We represent our features as X:

$$X_i = \phi_i(I), \quad i \in \{\text{face}, \text{body}\}, \tag{1}$$

where ϕ_i is the backbone for a given input type. Given that our framework aims to use facial information effectively, Fig. 2 displays the steps involved for gender recognition in frontal images. However, we are able to adapt our framework to consider any pose images through pose data processing, following our head detection approach. Representing θ as said approach, we define gender recognition of the nth input as:

$$\hat{y}_n = \begin{cases} f(\rho(X_{\text{face}_n}, X_{\text{body}_n})), & \text{if } \mathbb{1}_{\theta(I_n)=\text{Frontal}} \\ f(X_{\text{body}_n}), & \text{otherwise}, \end{cases} \tag{2}$$

where f is a sequence of *Average Pooling*, *FC*, and *Batch Normalization* layers, ρ is our Face Attention Module (FAM), $\mathbb{1}$ is the indicator function, and X are the features extracted from a given body part.

B. FACE ATTENTION MODULE

Having the body and face features extracted, we combine them efficiently by promoting the focus of YY-Net in the most relevant regions of the body. We design a weighted face-body feature combination by using a learnable weights matrix: $F \in \mathbb{R}^{H_i \times W_i \times C_i}$. We denote the features combination X_c as:

$$X_c = X_{\text{body}} \odot F \odot X_{\text{face}}, \tag{3}$$

where $\odot$ represents the Hadamard product, F is our fusion matrix and X are the extracted features. The importance of each body part is expected to vary depending on the image quality, person pose, and partial occlusions. For this reason, we add a channel-attention sub-network, based on the Squeeze-and-Excitation block [47], to modulate inter-channel dependencies. The combined input X_c passes through a series of linear and nonlinear layers, as shown in the *Classifier* part of Fig. 2, yielding a weight vector for the feature importance across channels. Then, we induce a residual link of this vector with body features to promote information complementary. We express the FAM output as:

$$\rho = \psi(X_c) + X_{\text{body}}, \tag{4}$$

where ψ is the sequence of linear and nonlinear layers previously described, X_c refers to the result of (3), and X_{body} corresponds to the extracted body features.

FIGURE 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.

C. IMPLEMENTATION DETAILS

Images are resized to 256×192 with random horizontal mirroring as inputs. The Stochastic Gradient Descent (SGD) algorithm 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 backbones and 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 32 and the number of training epochs is set to 30, using Binary Cross Entropy with Logits (BCELogits) as loss function.

IV. FRONTAL PAR AND FACE DATASETS

We create frontal and *wild* face versions of three well-known PAR datasets. The frontal images are obtained from a division of three possible values: *frontal*, *sideways*, and *backside*. We differentiate these values based on pose information from Alphapose [8]–[10]: if the rightmost shoulder, from the top left of the image, is the left shoulder, the subject faces forward (*frontal*); otherwise, the subject is facing backward (*backside*). Regarding *sideways* classification, we attribute this label to subjects with a shoulder length, with respect to the upper body height, less than 0.5. Examples of the pose classes considered are presented in Fig. 3.

A. HEAD DETECTION

Head ROI detection also uses pose information, cropping head regions from frontal images. Initially, we obtain the coordinates x and y of the right and left ears, considering the head's central point as the arithmetic mean of such coordinates. Then, head ROIs are drawn using the top-left and bottom-right bounding boxes coordinates, centered in the head's central point, with the head ROI height as $\frac{2}{9}$ of the whole body silhouette height. Head bounding boxes drawing is based on Detectron [48]. To ensure the versatility of our

FIGURE 4. Comparison of the head ROI detection results between MTCNN, EXTD, RetinaFace, and our approach, in images with varying quality, head pose and blurriness. Green boxes denote cases where the method accurately locates the face, and red boxes denote the opposite. Dotted red boxes refer to cases where the model over-predicted the number of existent faces. All images are resized to the same resolution for visualization purposes.

approach, head ROI detection is applied to the original image, before resizing.

Based on the pose data, this head ROI detection approach can detect the head of multiple people in an image. However, although various people may appear in PAR dataset images, pedestrian attributes are typically related to the main/central cropped person. Therefore, given this setting, this approach only detects the head of the main person by considering, for each image, the pose data outputted from Alphapose with the highest confidence. This allows it to retrieve head regions from frontal images with varying quality and/or partial occlusions. Other face detectors, such as MTCNN [49], EXTD [50], and RetinaFace [51], fail to detect head region as reliably as the used approach in the considered (wild) image conditions. To show the versatility of this method, Fig. 4 presents some examples of head ROI detection in frontal PAR images. In our head detection comparison, we use the default implementation of all methods.

We observe that MTCNN only detects faces in higher quality images, failing to detect in lower quality ones. This is expected since MTCNN was not designed to target "small faces" specifically, unlike EXTD and RetinaFace. Nonetheless, these methods fail to detect faces in low quality

TABLE 1. Dataset image-based feature analysis. Values are normalized for the combination of all dataset values, for each feature. Datasets named with *Frontal* suffix refer to datasets composed solely of frontal images.

Dataset	Resolution	Luminosity	Blurriness
PETA	0.037 ± 0.024	0.432 ± 0.100	0.120 ± 0.148
PETA$_{Frontal}$	0.037 ± 0.023	0.437 ± 0.101	0.140 ± 0.158
PA-100K	0.062 ± 0.063	0.449 ± 0.126	0.095 ± 0.089
PA-100K$_{Frontal}$	0.061 ± 0.064	0.459 ± 0.127	0.100 ± 0.090
RAP	0.131 ± 0.083	0.407 ± 0.107	0.022 ± 0.013
RAP$_{Frontal}$	0.138 ± 0.090	0.416 ± 0.108	0.023 ± 0.013

images reliably. EXTD was able to detect faces in more images but outputs more incorrect detections as well. RetinaFace accurately detects faces in more images than MTCNN but is not able to detect faces when quality is subpar, with a background color similar to the person (second and fourth column) or with blurry faces (third and fifth column). Our approach detects the facial region in all presented scenarios, regardless of the image quality. This is an important aspect, given that we intend to provide our model with facial images from surveillance scenarios, with varying image quality. Furthermore, we use the described approach to create *wild* face datasets, representing a challenging set for soft biometrics classification. The variability encountered in these datasets is displayed in Fig. 5.

B. DATA ANALYSIS

To prove that frontal datasets preserve the main characteristics of the original one, we analyze the major image-based factors that influence gender recognition. Note that train/test sets follow the set divisions of each PAR dataset (*i.e.*, we did not alter the train and test set division). The frontal datasets sets are images, from the original train/test sets, that match the defined frontal criterion (defined in IV). We consider *resolution*, *luminosity*, and *blurriness* as the main image-based features: *resolution* is retrieved from image width and height multiplication; for *luminosity*, we use the red, green, and blue value channels to measure the perceived brightness [52]; and *blurriness* yields from the convolution of the images with a Laplacian kernel, taking the variance as result. The image-based feature analysis of the used PAR datasets, and respective frontal versions, is shown in Table 1.

The obtained results suggest that frontal datasets preserve the characteristics of the original version. As such, they are not easier for gender recognition, and the accuracy variance of models on these datasets (when compared to the original ones) derives from pose and face access. Furthermore, given the resolution and blurriness values, RAP can be considered the dataset with the highest quality, while PETA is the lowest quality one.

Another conducted experiment was the evaluation of image number ratio between *test* and *all* (train + test) images. Furthermore, we assess the image number ratio of frontal and PAR datasets, for the test and train set, to verify if the pose division was balanced. We present the ratios analyzed

FIGURE 5. Examples of image quality variability in *wild* face datasets. The datasets contain subjects with varying facial expressions, levels of cooperativeness, and partial occlusions. Each row refers to one variability factor, and each group of four columns regards examples of a PAR dataset.

TABLE 2. Number of images, test/all image number ratio for each dataset, and frontal/PAR image number ratio, for train and test set. Frontal datasets are represented with the subscript term *Frontal* and test/all ratios are represented by dataset Ratio.

Dataset	Train		Test		Dataset Ratio
	Images	Ratio	Images	Ratio	
PETA	11400	0.379	7600	0.383	0.400
PETA_Frontal	4318		2918		0.403
PA-100K	90000	0.379	10000	0.354	0.100
PA-100K_Frontal	34128		3539		0.093
RAP	33268	0.339	8317	0.346	0.200
RAP_Frontal	11275		2880		0.203

TABLE 3. Gender mA of different models on frontal PAR datasets. The outperforming method for each dataset is shown in bold.

Methods	PETA_Frontal	PA-100K_Frontal	RAP_Frontal
ALM [38]	91.27	91.23	95.71
APR [27]	92.49	91.63	95.80
DeepMar [26]	91.09	90.75	95.96
VAC [39]	91.17	90.15	94.49
StrongBase [29]	92.62	92.05	96.14
YY-Net	**93.45**	**92.79**	**97.07**

in Table 2. The test/all images ratios are very similar between all PAR datasets and their frontal versions. Additionally, frontal/PAR ratios are similar between datasets and around $\frac{1}{3}$, which is expected given that the dataset pose division is based on three values (*frontal*, *sideways*, and *backside*).

Regarding the *wild* face datasets, all the ratios presented for frontal datasets are analogous for the face ones. The resolution, luminosity, and blurriness of face datasets are similar to the frontal dataset values, given that the facial regions were retrieved from frontal images.

V. EXPERIMENTS

A. DATASETS, METHODS, AND EVALUATION METRICS

Our experiments are carried out in three state-of-the-art PAR datasets: PETA [20], PA-100K [21], and RAP [22]. The PETA dataset has 19,000 images, divided into three subsets: 9,500 for training, 1,900 for validation, and 7,600 for testing. 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. The RAP dataset contains 41,585 images. Following the official protocol [22], we split the dataset into 33,268 training images and 8,317 test images. Furthermore, we carried out experiments in the proposed frontal datasets, consisting of frontal images of the described PAR datasets.

The PAR methods used in our experiments are VAC [39], APR [27], ALM [38], StrongBase [29], and DeepMar [26]. All models are evaluated using the original implementations

with minor adaptations to only classify gender and not all pedestrian attributes.

To present our results, we use the label-based metric mean Accuracy (mA) [22], which is the mean classification accuracy of the positive and negative samples for each attribute, averaged over all attributes:

$$mA = \frac{1}{2n} \sum_{i=1}^{m} \left(\frac{tn_i}{p_i} + \frac{tp_i}{n_i} \right), \tag{5}$$

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 ith attribute; tn_i and n_i are defined analogously. We adapt this metric to only consider the gender attribute, from the PAR datasets.

B. GENDER RECOGNITION

Given that our focus is to improve wild gender recognition, combining facial and body information, we start by assessing the performance of YY-Net in frontal PAR datasets. We compare it with five state-of-the-art PAR methods, presenting our results in Table 3.

Our approach is better than all considered methods, which corroborates the importance of effectively combining face and body images in gender classification. When comparing with the best result for each dataset, we improve gender accuracy by 0.83%, 0.74%, and 0.93% in PETA, PA-100K, and RAP, respectively. The improvement in all datasets demonstrates the versatility of YY-Net, corresponding to an error reduction of 11.25%, 9.31%, and 24.09%, respectively.

TABLE 4. Gender mA of different models on PAR datasets. The outperforming method for each dataset is shown in bold.

Methods	PETA	PA-100K	RAP
ALM [38]	92.28	90.34	95.69
APR [27]	92.84	90.05	96.20
DeepMar [26]	92.33	90.39	96.40
VAC [39]	92.85	91.05	96.59
StrongBase [29]	93.13	90.77	96.74
YY-Net	**93.39**	**91.20**	**96.86**

TABLE 5. Gender mA of YY-Net in frontal datasets, when trained in PAR datasets and in its frontal version. Frontal datasets are represented with the subscript term *Frontal*. The highest evaluation accuracy for each dataset is shown in bold.

Train \ Test	$PETA_{Frontal}$	$PA\text{-}100K_{Frontal}$	$RAP_{Frontal}$
PETA	**93.88**	78.67	81.47
$PETA_{Frontal}$	93.45	77.38	78.09
PA-100K	79.02	92.70	89.51
$PA\text{-}100K_{Frontal}$	77.98	**92.79**	89.10
RAP	73.06	75.16	96.95
$RAP_{Frontal}$	73.51	77.94	**97.07**

The increased error reduction displayed in RAP could be linked to its higher image quality, relative to the other PAR datasets. High quality images might benefit more from effective facial and body information conjunction, giving the increased body details, culminating in improved relevant body portion focus.

With the intent to analyze the adaptability of our model to deal with unconstrained pose scenarios, we evaluate its performance in PAR datasets. We present, in Table 4, the accuracy of the analyzed state-of-the-art PAR methods, for each dataset.

Our results show that the combination of face and body information, even in the context of unconstrained poses, contributes to slight improvements. This effect was observed for all three datasets. Compared to the second best method, YY-Net improved by 0.26%, 0.15%, and 0.12% in PETA, PA-100K, and RAP, respectively, which corresponds to an error reduction of 3.79%, 1.68%, and 3.68%. The results suggest that YY-Net is reliable in frontal pose images and applicable in wilder contexts, with varying poses.

C. WITHIN AND CROSS-DOMAIN PERFORMANCE

To assess if the performance improvements in PAR datasets derive essentially from more accurate results for frontal images, we evaluate YY-Nets in frontal datasets, using the weights from PAR training and compare it to YY-Nets trained solely on frontal images. We report our results in Table 5, exploring *within* and *cross-domain* settings.

This experiment shows that models trained on different datasets have distinct performances, linked to dataset length and quality differences. If we analyze the *within-domain*, only PETA has a significant difference between PAR and frontal training models, with the PAR model obtaining
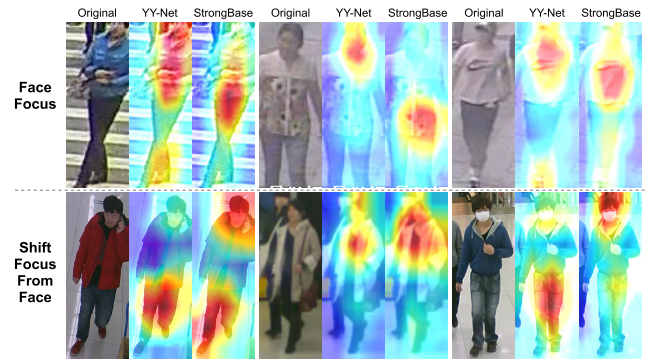
FIGURE 6. Comparison of StrongBase and YY-Net attention focus from images where YY-Net correctly classifies gender and StrongBase does not. Each row is related to one of the categories of diverging focus.

better performance. Models trained in PAR datasets do not interact with more frontal images than those trained in frontal datasets. However, they are trained in more images and with more varying body poses. This contributes to better body information processing which, in conjunction with our face combination approach, translates into a more accurate gender recognition. In this scenario, PETA training is more beneficial than only $PETA_{Frontal}$, given that it is the dataset with the lowest number of training images. Regarding the RAP and PA-100K, their frontal versions contain enough training images that PAR training does not translate into increased *within-domain* performance.

When we analyze the *cross-domain* performance, we observe that PAR training is more beneficial to PETA and PA-100K, with RAP having better performance with frontal training only. These differences could be associated with image quality, where RAP is the highest quality dataset, given our results in Table 1. This is particularly relevant, given that high-quality images have increased face importance, particularly in frontal images. As such, training solely in frontal images may promote a more effective face data processing, translating into a better face-body combination. In this case, training for various poses might dissipate this aspect leading to a poorer generalization, expressed by worse *cross-domain* performance in RAP. For PETA and PA-100K, more training data (PAR) promotes better generalization, as the *cross-domain* results demonstrate.

D. FACE ATTENTION INFLUENCE

Given that YY-Net uses face information to aid in gender recognition, we assess whether this approach would, implicitly, focus more on faces to perform its task. To achieve our goal, we compare Class Activation Maps (CAM) [53] from our model with the StrongBase's ones, from images where StrongBase exclusively misclassifies. This is the chosen method since it outperformed the remaining ones. We group all three datasets and present the CAM of representative images in Fig. 6.

The displayed examples suggest that our FAM does not force a predominant face focus but does contribute to
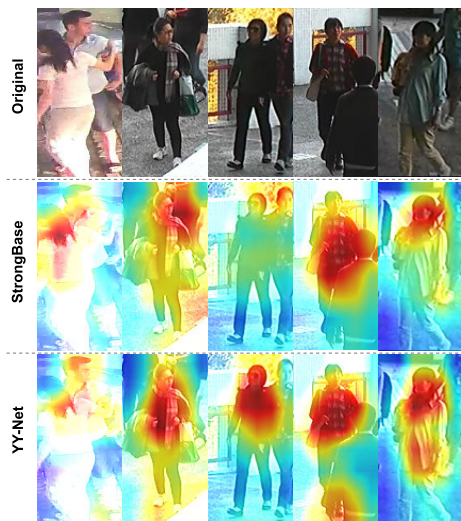
FIGURE 7. Crowd effect on attention focus from StrongBase and YY-Net. Each column displays an example where YY-Net accurately classifies gender and StrongBase misclassifies it.

TABLE 6. Performance comparison, on PETA$_{Frontal}$ dataset, when gradually adding each component to the baseline model. Variants of the same key parts lie in the same group. Components in bold are the adopted approach in YY-Net.

Component	mA
Baseline (Face Input)	91.13
Baseline (Body Input)	92.49
Face-Body Combination (FBC)	92.65
Weighted FBC (WFBC)	**93.10**
WFBC + Body	93.26
Attention + WFBC	93.31
Attention + WFBC + Body (YY-Net)	**93.45**

a more coherent face-body combination. In the first row (*Face Focus*), we observe a hip/upper body focus from StrongBase. In these cases, the facial region is a more discriminative factor (relative to the hip/upper body region), which is the focus of YY-Net. However, there are instances where the facial region may difficult/misguide gender recognition. In the second row, we observe cases where the facial region is not easily disclosed or presents unisex qualities. In these cases, analyzing other body portions may be beneficial, which is the approach taken by our model. StrongBase opts to focus more on the neck/facial region, contributing to gender misclassification. Another scenario where YY-Net has better performance is in crowded situations, where it is able to focus on one subject, while StrongBase underperforms in the same context. We display images with multiple subjects where our model accurately classifies gender and StrongBase does not, in Fig. 7.

Influenced by FAM, YY-Net is capable of focusing on the main subject in the image. By contrast, StrongBase has performance issues when other people appear in the front/background, dispersing its attention to other subjects in the image. This translates into a more scatter focus on the main subject, contributing to incorrect gender classification. In the image of the rightmost column, YY-Net complements facial region information with hip one, contrasting with the StrongBase, where the facial region was the main (sole) focus. The approach of YY-Net is justifiable given the background ''noise'' of other people, namely an arm near the facial region, and the androgenic qualities of the subject.

The analysis of Figs. 6 and 7 focuses on the model attention reasoning regarding its face/body balance importance and crowded scenarios, respectively. Note, however, that said reasonings might not be extensible to all dataset images. To further compare YY-Net and StrongBase gender classification performance, we process the output of each model and assess the difference between each prediction and the ground truth label. We perform this comparison in a set of RAP and PA-100K images, where label 0 refers to male and 1 to female, and create a scatter plot, displayed in Fig. 8. We focus our analysis on four different regions, representing cases where a model outperformed the other by the largest margins. Furthermore, we display representative images of each region, which will be the basis of our analysis.

In the cases where YY-Net outperforms StrongBase, our model efficiently distinguishes females and males with androgenic/unisex appearance. For both genders, the pose, clothing, and face appearance misguides StrongBase in its classification, with YY-Net being more resilient. If we examine the cases where StrongBase was the outperforming model, we observe that shorter hair females, with unisex clothing, are situations where the performance of our model is subpar. Additionally, males with faces blurred, partial occluded, or with slightly uncooperative (misguided) poses might represent a challenge for YY-Net. This suggests that, although our model can handle subjects with androgenic/unisex qualities, these are also troublesome scenarios, particularly for females. Based on the typical appearance of females in PAR datasets (longer hair without unisex clothing), YY-Net might have prioritized the facial region over body information, thus contributing to gender misclassification. For males, and females to a lesser extent, face blurriness might be a major contributing factor for misclassification. Given our subjacent approach of combining face with body information to classify gender, this could justify the underperformance in the stated conditions.

E. ABLATION STUDIES
To validate the sensitivity of the obtained results with respect to the major components of YY-Net, we evaluate the influence of each one on gender recognition accuracy. As *Baseline*, a ResNet50 network with the modifications described in Section III-A and a classifier without the FAM are used. Furthermore, we analyze the influence of increment/variance of the components used for two key parts: 1) combination of face and body information; and 2) focus in body part attention mechanisms. We present our results in Table 6.
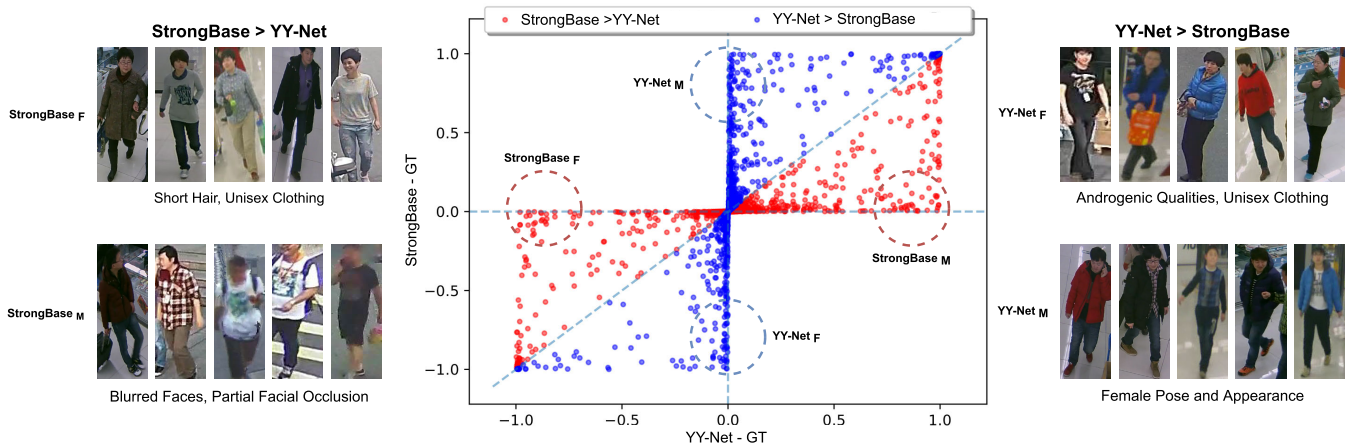
FIGURE 8. Scatter plot of the difference between YY-Net and StrongBase predictions, and ground truth (GT) labels, in RAP and PA-100K images. Points in blue represent cases where YY-Net outperforms StrongBase, and points in red denote the opposite. Four different regions are highlighted in the plot, displaying the cases where a model outperforms the other the most, for male (M) and female (F) recognition. Representative images, from the four regions, are presented sideways of the plot, accompanied by descriptive labels.

To obtain an initial baseline, we start by evaluating the importance of YY-Net components by assessing gender recognition using only face or full-body images. The results obtained by both baselines denote that full-body images (which also contain the facial region) promote a higher gender recognition accuracy than facial images. Nevertheless, neither facial nor body information by themselves are enough to obtain YY-Net performance.

We assess the effect of combining face and body information through the Hadamard product, denoted as *Face-Body Combination* (FBC) in Table 6. This approach translates into higher mA than the baseline with body input, corroborating the idea that both inputs are valuable for gender recognition in frontal images. However, this was only a slight improvement, suggesting that the combination of the input is inefficient. Adding the learnable fusion matrix *F* to influence facial and body information combination, denominated as *Weighted FBC* in the Table, promotes a mA increase.

The evaluation of the attention approach (*Attention*) and the preservation of complementary information (*Weighted FBC + Body*) is also a focus of our experiments. *Attention* refers to the sequence of linear and nonlinear layers (described in Section III-B) to modulate inter-channel dependencies. The residual approach of complementing the combined face-body information with body input contributes to higher gender recognition accuracy. This notion is linked to contextualizing face-body information, while using all the available information (full-body). In practice, this strategy intends to deviate the model from heavy facial region focus, regardless of the subject appearance. The attention method, along with the weighted face-body combination, provide an efficient focus on the critical parts of the body to classify gender, obtaining a similar performance to only using complementary information. Finally, *YY-Net* combines the weighted FBC with attention, while also using complementary body information, which translates into an aggregated

improvement when compared with all other evaluated components. This approach promotes a balanced equilibrium between face and body information, while focusing on the most influential body portions for gender recognition.

VI. CONCLUSION

Subjects pose is known to hinder gender recognition in wild conditions, influencing the focus of models on different subject features. In this context, frontal images typically bias the focus towards the facial region, while the body silhouette is the main focus in non-frontal data. Given that the state-of-the-art usually evaluates gender in good quality face datasets, we present frontal and *wild* face versions of well-known PAR datasets, created through pose data processing. These subsets supply the learning data to our proposed model, YY-Net, designed to effectively complement face and body information, making it suitable for gender classification *in-the-wild*. In our proposal, the key is to use a learnable fusion matrix and channel-attention sub-network, which enables a dynamic focus on the most relevant body regions relative to specific image/subject features, achieving state-of-the-art results in PAR datasets.

The announced versions of the *wild* face datasets differ from the existing ones, presenting a more challenging set to classify soft biometrics. Furthermore, the robustness displayed by YY-Net to classify gender in wild conditions supports its usability as a basis for subsequent developments in soft biometrics analysis. Extending our approach to be applied in the context of pedestrian attribute classification in video settings [54] or classifying attributes of multiple people in a given scene are potential improvements for future work. Finally, given the limitations of YY-Net for specific group cases, the implementation of an ensemble model based on holistic gender analysis (StrongBase) and face-body balance (YY-Net) is also a direction for future work.

REFERENCES

[1] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Proc. Int. Conf. Biometric Authentication.* Cham, Switzerland: 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 Trans. Inf. Forensics Security*, vol. 11, no. 3, pp. 441–467, Mar. 2016.

[3] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[4] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, pp. 144–157, Apr. 2018.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[6] S. Escalera, M. A. Bagheri, M. Valstar, M. T. Torres, B. Martinez, X. Baro, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, and M. Oliu, "ChaLearn looking at people and faces of the world: Face AnalysisWorkshop and challenge 2016," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1–8.

[7] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Sep. 2014.

[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[9] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10863–10872.

[10] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.

[11] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 34–42.

[12] 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. 22492–22503, 2017.

[13] H. Han, A. K. Jain, X. Chen, F. Wang, and S. Shan, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018.

[14] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25. Stateline, NV, USA, Dec. 2012, pp. 1097–1105.

[16] A. Dantcheva and F. Brémond, "Gender estimation based on smile-dynamics," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 3, pp. 719–729, Mar. 2017.

[17] W. Samek, A. Binder, S. Lapuschkin, and K.-R. M'uller, "Understanding and comparing deep neural networks for age and gender classification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1629–1638.

[18] D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "Multimodal age and gender classification using ear and profile face images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2414–2421.

[19] H. J. Ryu, H. Adam, and M. Mitchell, "InclusiveFaceNet: Improving face attribute detection with race and gender diversity," 2017, *arXiv:1712.00193*.

[20] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[21] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[22] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," 2016, *arXiv:1603.07054*.

[23] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 531–540.

[24] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3177–3183.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[26] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 111–115.

[27] 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 Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[28] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu, "Hierarchical feature embedding for attribute recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13055–13064.

[29] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," 2020, *arXiv:2005.11909*.

[30] T. Roxo and H. Proença, "Is gender 'in-the-wild' inference really a solved problem?" *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 4, pp. 573–582, Oct. 2021.

[31] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2018, pp. 1–6.

[32] J. Zhang, P. Ren, and J. Li, "Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints," 2020, *arXiv:2011.06798*.

[33] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–13.

[34] E. Yaghoubi, P. Alirezazadeh, E. Assunç ao, J. C. Neves, and H. Proença, "Region-based CNNs for pedestrian gender recognition in visual surveillance environments," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2019, pp. 1–5.

[35] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregon, and M. Castrillon-Santana, "Gender classification on 2D human skeleton," in *Proc. 3rd Int. Conf. Bio-Eng. Smart Technol. (BioSMART)*, Apr. 2019, pp. 1–4.

[36] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 680–697.

[37] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. González, "Age and gender recognition in the wild with deep attention," *Pattern Recognit.*, vol. 72, pp. 563–571, Dec. 2017.

[38] C. Tang, L. Sheng, Z.-X. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4997–5006.

[39] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 729–739.

[40] Y. Li, H. Xu, M. Bian, and J. Xiao, "Attention based CNN-ConvLSTM for pedestrian attribute recognition," *Sensors*, vol. 20, no. 3, p. 811, Feb. 2020.

[41] H. Fan, H.-M. Hu, S. Liu, W. Lu, and S. Pu, "Correlation graph convolutional network for pedestrian attribute recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 49–60, 2022.

[42] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and C. Yan, "Recurrent attention model for pedestrian attribute recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, Jul. 2019, pp. 9275–9282.

[43] Z. Ji, Z. Hu, E. He, J. Han, and Y. Pang, "Pedestrian attribute recognition based on multiple time steps attention," *Pattern Recognit. Lett.*, vol. 138, pp. 170–176, Oct. 2020.

[44] L. Gao, D. Huang, Y. Guo, and Y. Wang, "Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1340–1348.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[48] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2021). *Detectron2*. Accessed: Nov. 15, 2021. [Online]. Available: https://github.com/facebookresearch/detectron2

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[50] Y. Yoo, D. Han, and S. Yun, "EXTD: Extremely tiny face detector via iterative filter reuse," 2019, *arXiv:1906.06579*.

[51] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.

[52] A. Morgand and M. Tamaazousti, "Generic and real-time detection of specular reflections in images," in *Proc. VISAPP 9th Int. Conf. Comput. Vis. Theory Appl.*, vol. 1, Jan. 2014, pp. 274–282.

[53] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[54] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gotcha-I: A multiview human videos dataset," in *Security in Computing and Communications*, S. M. Thampi, G. M. Perez, R. Ko, and D. B. Rawat, Eds. Singapore: Springer, 2020, pp. 213–224.

TIAGO ROXO (Member, IEEE) received the bachelor's degree in computer science and engineering from the Universidade da Beira Interior (UBI), in 2019. He is currently pursuing the Ph.D. degree in the field of computer vision and artificial intelligence with Fundação para a Ciência e a Tecnologia (FCT) Scholarship.

HUGO PROENÇA (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Beira Interior. He has been researching mainly about biometrics and visual-surveillance. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and an Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM journal. He is a member of the editorial boards of the *Image and Vision Computing*, IEEE ACCESS, and *International Journal of Biometrics*. Also, he served as a Guest Editor of Special Issues of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image and Video Processing* journals.

• • •