IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 1, 2020, accepted October 10, 2020, date of publication October 13, 2020, date of current version October 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030859

EPYNET: Efficient Pyramidal Network for Clothing Segmentation

ANDREI DE SOUZA INÁCIO[1,2] AND HEITOR SILVÉRIO LOPES[1]

[1]Graduate Program in Electrical Engineering and Industrial Informatics, Federal University of Technology – Paraná, Curitiba 80230-901, Brazil
[2]Federal Institute of Santa Catarina, Gaspar 89111-009, Brazil

Corresponding author: Andrei de Souza Inácio (andrei.inacioo@gmail.com)

The work of Heitor Silvério Lopes was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grants 311785/2019-0, 311778/2016-0, and 423872/2016-8; and in part by the Fundação Araucária under Grant PRONEX 042/2018.

ABSTRACT Soft biometrics traits extracted from a human body, including the type of clothes, hair color, and accessories, are useful information used for people tracking and identification. Semantic segmentation of these traits from images is still a challenge for researchers because of the huge variety of clothing styles, layering, shapes, and colors. To tackle these issues, we proposed EPYNET, a framework for clothing segmentation. EPYNET is based on the Single Shot MultiBox Detector (SSD) and the Feature Pyramid Network (FPN) with the EfficientNet model as the backbone. The framework also integrates data augmentation methods and noise reduction techniques to increase the accuracy of the segmentation. We also propose a new dataset named UTFPR-SBD3, consisting of 4,500 manually annotated images into 18 classes of objects, plus the background. Unlike available public datasets with imbalanced class distributions, the UTFPR-SBD3 has, at least, 100 instances per class to minimize the training difficulty of deep learning models. We introduced a new measure of dataset imbalance, motivated by the difficulty in comparing different datasets for clothing segmentation. With such a measure, it is possible to detect the influence of the background, classes with small items, or classes with a too high or too low number of instances. Experimental results on UTFPR-SBD3 show the effectiveness of EPYNET, outperforming the state-of-art methods for clothing segmentation on public datasets. Based on these results, we believe that the proposed approach can be potentially useful for many real-world applications related to soft biometrics, people surveillance, image description, clothes recommendation, and others.

INDEX TERMS Soft biometrics, clothing segmentation, computer vision, deep learning.

I. INTRODUCTION

Soft biometrics is an emerging area of research, mainly due to its extensive applicability in human identification and surveillance. Human attributes such as gender, age, hairstyle, tattoos, as well as type and color of clothes, are examples of soft biometrics traits that can be successfully used for distinguishing different people [1].

In the last years, physiological and behavioral human characteristics have been used for the human identification task [2]. However, identifying human beings accurately in a real visual surveillance system is still challenging, mainly due to the low-resolution of cameras, poor illumination, and the diversity of camera angles [3].

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

Unlike traditional biometrics approaches, such as fingerprint and iris, soft biometrics extraction is non-invasive and requires no contact. Therefore, it can be done at distance and, more importantly, without the explicit cooperation of the targets [4].

Many soft biometrics approaches have explored handcrafted features on datasets collected under well-constrained environments [5], [6]. In recent years, Deep Learning (DL) methods have been successfully used to learn compact and discriminative features for many complex problems. In the soft biometrics area, DL methods have been used for a wide variety of classification problems, such as body-based features (height, shoulder width, hips-width, arms-length, body complexion, and hair color) [7], facial traits (gender, ethnicity, age, glasses, beard, and mustache) [8], [9], tattoos [10], gait-based gender [11], and clothes [12]–[14].

The segmentation of clothes in images, also known as clothing parsing [15], consists of classifying each image pixel with labels specifically related to clothes (and accessories). Currently, it is still a challenging topic and has aroused the interest among researchers due to the inexhaustible variety of types, styles, colors, and shapes of clothes [16].

Approaches developed to localize, classify, and segment clothing-related traits are essential to several tasks, including the identification of a person in surveillance videos, semantic enrichment of image description, and overall fashion analysis [17]. Although many efforts have been done to develop models to perform the task of clothing segmentation, such approaches appear to be unsatisfactory for real-world applications.

For tackling the clothing segmentation problem in the context of soft biometrics, we propose a new framework named EPYNET. The framework is focused on human attributes, such as skin, hair, type of clothes, and accessories. The proposed approach uses the SSD (Single Shot MultiBox Detector) [18] model to crop a person from the image, and the Feature Pyramid Network (FPN) architecture [19] with the EfficientNet [20] model to perform the segmentation task. Data augmentation techniques and noise reduction were also applied to improve the performance of the method.

It is known that DL methods usually require large amounts of data to be trained. For the specific problem of clothing segmentation, there are some popular datasets, for instance: CFPD [21] and Fashionista [22]. However, several incorrect labeling can be found in the CFPD dataset, since annotated images are based on superpixels. Also, the Fashionista dataset is quite small for DL methods, since it contains only 685 annotated images. Therefore, the need for a new, high-quality, benchmark dataset, led to the creation of a new one, containing 4,500 images manually annotated into 18 classes, besides the background. This paper is an extension of a preliminary work presented in [23] and differs from it in four points: (i) we propose a new dataset, called UTFPR-SBD3,[1] with improved quality over existing datasets, focused in the soft biometrics context; (ii) a new measure, named Instances-Pixels Balance Index – IPBI, is proposed to compare the balance of different datasets in terms of pixels and instances; (iii) we propose a novel framework based on FPN and EfficientNet, which can extract high-quality features at different spatial resolutions; (iv) extensive experiments and comparisons were done with public benchmarks for clothing segmentation. To the best of our knowledge, all the mentioned contributions are new and not published anywhere.

The remainder of this paper is organized as follows: Section II presents a brief description of related works. In Section III, we present a new dataset, named UTFPR-SBD3, for clothing segmentation, a new measure of instances and pixels balance for image segmentation datasets, and the proposed framework for clothing segmentation. The experimental results and their discussion are shown in Section IV. Finally, general conclusions and suggestions for future research directions are presented in Section V.

[1]The dataset is available at http://labic.utfpr.edu.br/datasets/UTFPR-SBD3.html

II. RELATED WORKS

DL methods have been utilized to address many problems in the surveillance environment, including person re-identification [24], face recognition [9], anomaly detection [25], multi-view analysis [26], and traffic monitoring [27].

In the soft biometrics area, several approaches have been proposed to localize, segment, and classify clothes in digital images for many application scenarios, including outfit retrieval, fashion analysis, surveillance, and human identification.

In [12], the authors proposed an approach to classify soft biometrics traits using Convolutional Neural Networks (CNN). Authors used independent classifiers to detect the gender (male or female), upper clothes (short or long sleeves), and lower clothes (short or pants) of a person. Although their approach achieved remarkable generalization capability of the model, they reported difficulties in finding appropriate image datasets, regarding size, quality, and variability.

A method for clothing segmentation using an adaptation of the U-Net architecture was presented in [28]. This model was adapted to accommodate multi-class segmentation and was trained with the Clothing Co-Parsing (CCP) dataset [29]. Due to the significant similarity between classes with few instances, 58 different classes were grouped into 14. The authors concluded that the U-Net model could be a reliable way to perform the segmentation task, Authors also pointed out the lack of large datasets annotated at the pixel level.

An approach based on a Fully-Convolutional Network (FCN) for the clothing segmentation task was proposed by [30]. The proposed architecture extends the FCN with a side path called ''outfit encoder'' to filter inappropriate clothing combination from segmentation, and a post-processing step using Conditional Random Field (CRF) to assign a visually consistent set of clothing labels. These authors introduced a refined annotation of the Fashionista dataset with 25 classes to study the influence of erroneous annotations and ambiguous classes on performance metrics.

[14] also proposed an FCN-based approach to compute the color and the class of pixels. First, a Faster-RCNN (Recurrent CNN) model is used to detect and crop people in the image. Then, the cropped image of a person is used to feed the FCN, which computes color and class feature maps. Finally, a logistic pooling layer combines these features, and then the color e class is predicted.

More recently, [31] proposed the superpixels features extractor network (SP-FEN). The proposed model is based on the FCN with the introduction of superpixels encoder as an aside-network that feeds the extracted features into the main segmentation pipeline. Data augmentation techniques such as flip, rotation, and deformation were used during the training step to improve generalization performance. The

authors in [32] propose the ResNeXt-FPN approach that also uses the FPN architecture with ResNeXt to perform semantic segmentation of clothes. In this work, the authors concluded the use of a FPN based-approach is feasible for clothing segmentation.

In general, the use of DL has shown promising results for the clothing segmentation task. However, as pointed out before, a common challenging problem reported by several authors is the difficulty to distinguishing between clothes used in the same body part, for example t-shirt and blouse or hair and hat. Another challenging reported is the lack of annotated datasets required for training DL models. Most of the available datasets have highly imbalanced classes (classes with less than 20 instances), inconsistent and noisy annotations, and high similarity between classes. In Section III-A a new dataset will be presented, aiming at overcoming these problems.

III. METHODS

This section first presents the proposed dataset for clothing segmentation, named UTFPR-SBD3, and a dataset comparison method. In the sequence, the proposed method for clothing segmentation called EPYNET is presented. Finally, the evaluation procedure of the method is shown.

A. UTFPR-SBD3 DATASET

Previous works in the literature have proposed datasets for clothing segmentation and classification. They vary in the number of images and clothing categories. Also, some include non-fashion classes, such as hair, skin, face, background, etc. To date, the most popular datasets are presented in Table 1.

TABLE 1. Comparison of popular datasets for clothing segmentation and the new UTFPR-SBD3.

Reference	Dataset	Images	Classes	Non-fashion classes
[29]	CCP	1,004	59	Yes
[21]	CFPD	2,682	23	Yes
[22]	Fashionista	685	56	Yes
[30]	Refined Fashionista	685	25	Yes
[33]	Modanet	55,176	13	No
(this work)	**UTFPR-SBD3**	**4,500**	**19**	**Yes**

The Clothing Co-Parsing (CCP) dataset contains 2,098 high-resolution fashion photos, of which only 1,004 were annotated at the pixel level. It is a highly imbalanced dataset, including classes without instances and many ambiguous classes (e.g., it has about 10 different types of footwear).

The Colorful Fashion Parsing (CFPD) dataset contains 2,682 images annotated with classes (23 different types, including the background) and colors (13 types). The dataset was annotated using a superpixel method, and it has a significant amount of noise, as shown in Figure 1b). It is known that the quality of training data in DL methods

FIGURE 1. Examples of annotation errors in the CFPD dataset. Figures in the left side present annotations with noise. The right side shows images with misannotation and an image with two similar subjects.

is essential for achieving a reasonable accuracy of the model. These observed noises, associated with high-class imbalance, can drastically affect the segmentation accuracy. Figure 1(a) and 1(c) shows the categories belts and sunglasses with partial annotation. Other annotation problems in the CFPD dataset are: items without annotations (e.g., footwear class in Figure 1(c) and 1(e)), incorrectly annotations (e.g., coat annotated as a skirt in Figure 1(b) and as a scarf in Figure 1(e)), and images with two or more subjects and only one annotation, as depicted in Figure 1(f).

Another dataset used for clothing segmentation is the Fashionista dataset. It consists of 685 images with pixel-level annotations in 56 different classes. To overcome the ambiguous categories of clothes presented in the dataset, [22] proposed the Refined Fashionista dataset by merging some categories (e.g., blazer and jacket). However, both datasets still have small images, as well as imbalanced classes.

Modanet contains 55,176 street fashion photos annotated with polygons in 13 classes. Despite being the largest dataset in the number of images, it was not considered in this study because it has only 13 classes, mainly for real-world commercial applications. Moreover, it does not include useful classes for soft biometrics contexts such as skin, hair, glasses, socks, or neck adornments.

In addition to the above-mentioned problems of the datasets, such as high imbalanced classes, ambiguous labels, and wrong annotations, in the context of soft biometrics, it is strongly desirable a dataset with human attributes, including hair and skin, so one could distinguish two individuals.

TABLE 2. Number of images and pixels for each class in UTFPR-SBD3.

#	Class	N. Pixel	N. Instance
1	Background	847864558	4500
2	Bag	13278257	2650
3	Belt	1332940	1364
4	Coat	25986090	1551
5	Dress	21684176	1065
6	Eyewear	775040	1349
7	Footwear	10735404	4448
8	Hair	18515544	4446
9	Headwear	1496376	782
10	Neckwear	2044846	383
11	Pants	23369711	1464
12	Romper/Jumpsuit	2491602	107
13	Shirt	27931674	3014
14	Shorts	4927357	747
15	Skin	47923010	4500
16	Skirt	17490640	1231
17	Socks	1182418	435
18	Stocking	3757929	450
19	Sweater	7212428	566

To overcome the drawbacks of the existing datasets, we constructed a new one, named UTFPR-SBD3 (Soft Biometrics Dataset #3), intended for clothing segmentation in the context of soft biometrics. It consists of 4,500 images manually annotated into 18 classes plus the background (see Table 2), of which 1,003 were taken from the CCP dataset, 2,679 from the CFPD, and 685 from the Fashionista dataset. Additionally, 133 more images containing instances of the less frequent classes in the dataset were collected from the fashion sharing website chictopia.com (same source of the CFPD and Fashionista datasets) to ensure that each class has, at least, 100 instances. All images were standardized to 400 × 600 pixels in RGB color. Figure 2 shows the classes distribution diagram of the datasets. These diagrams graphically expose the imbalance of classes present in the datasets.

To guarantee the high-quality of the dataset, all images were manually annotated at pixel-level using the JS Segment Annotator,[2] a free web-based image annotation tool. Raw images were carefully selected to avoid, as far as possible, classes with a low number of instances. However, it is important to notice that two classes, background, and skin, are naturally present in all images and, unfortunately, they occupy a significant part of the image. Later in this paper, such information will be important for the evaluate the comparative performance of the proposed method.

1) DATASET COMPARISON METHOD

In Table 1, the number of images and classes of the datasets frequently used for clothing segmentation are shown. Ideally, in a given dataset, the number of samples per class should be the same, so that no classifier would be biased towards the majority class. Since we are dealing with images of people dressed in clothes, not only clothes alone, it is quite difficult,

[2]https://github.com/kyamagu/js-segment-annotator

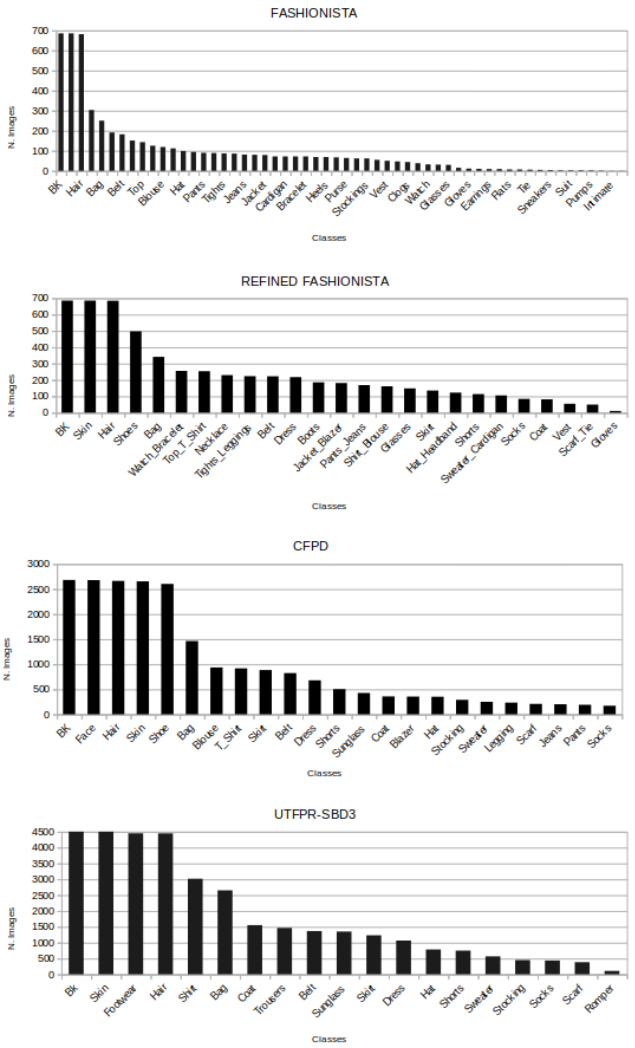

FIGURE 2. Distribution diagram of UTFPR-SBD3 dataset comparing the prior datasets.

if not impossible, to achieve a perfect balance between the several classes of clothes in a dataset. Furthermore, considering that the segmentation of the objects (clothes) is accomplished at the pixel level, the number of pixels for each class must be taken into account. As a matter of fact, different pieces of clothes (as well as non-clothes classes: skin, hair, and the background) may have quite different sizes. For instance, a bow tie or a belt is usually smaller than pants or overcoats. Therefore, the image segmentation problem we have in hand is naturally unbalanced.

Another important issue is how to deal with the background class. Some published works, when reporting the overall segmentation performance of the proposed methods, include the background class together with all other clothing classes. Keeping in mind that the main purpose of the applications in this area is to identify clothes, not the background, if considering the background as a class may lead to distorted results since it is the majority class in most images (see Section IV-E for more details).

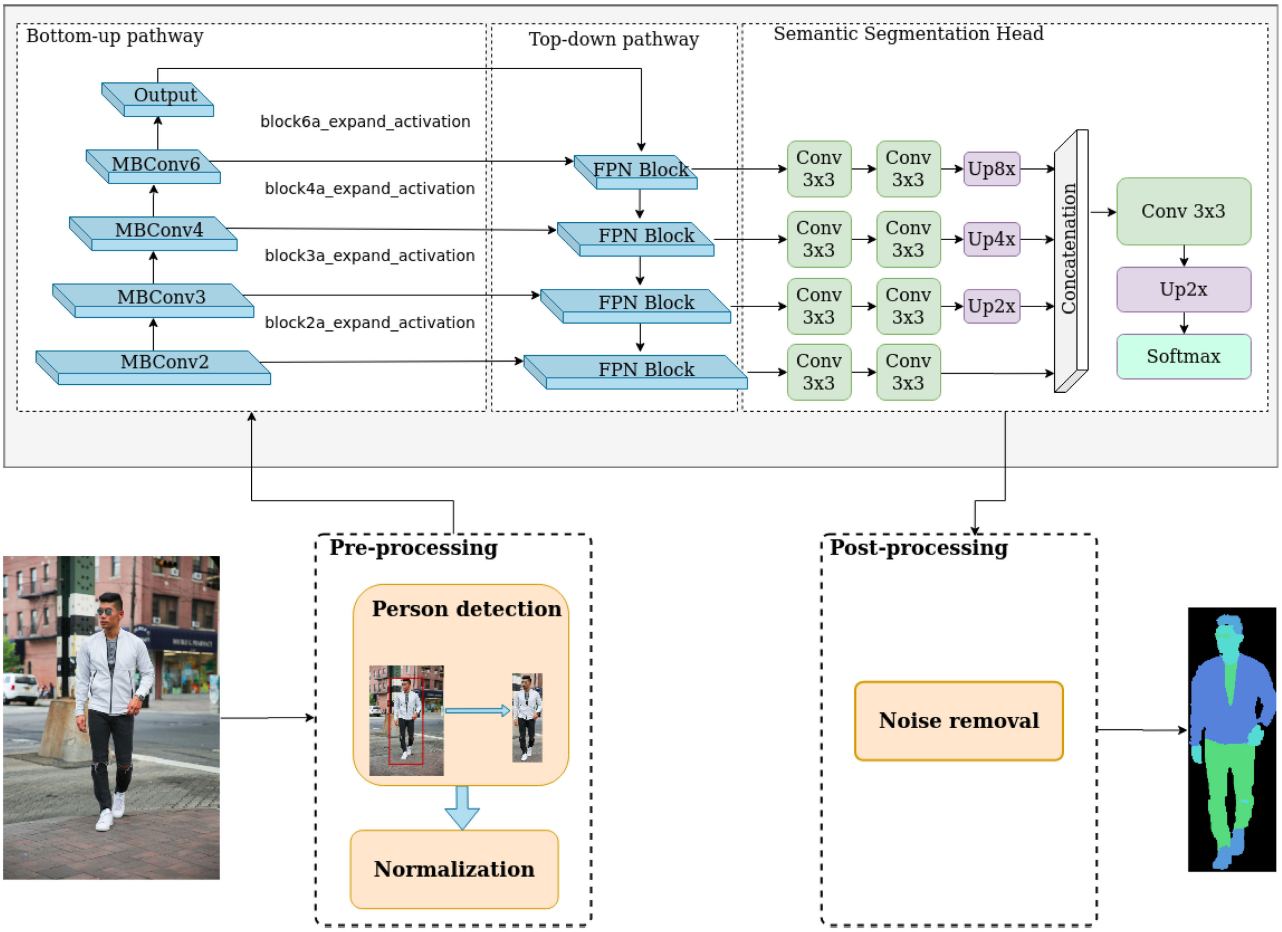

FIGURE 3. Overview of the proposed approach. Given an input image, the pre-processing step detects and crops the person found in the image. Then, the EPYNET performs the segmentation task. Finally, in the post-processing step, noise is removed, and the final predicted label is presented.

As a consequence, it is a difficult task to compare the popular datasets presented before with the proposed UTFPR-SBD3. Any measure of comparison should take into account both, the distribution of instances over classes and the number of pixels per instance. To meet such requirements, in this work, we propose the Instances-Pixels Balance Index – IPBI to compare the joint balance of instances and pixels of different datasets.

The IPBI is based on the concept of entropy, a common measure used in many fields of science, for which there are several definitions, depending upon the area. In a general sense, it measures the amount of disorder of a system. As mentioned before, for the sake of this work, the ideal dataset should have the same number of instances per class, as well as the same number of pixels in all classes.

For a dataset with c classes (labels), such that the i-th class has s_i instances (samples), and a total of k instances, the Shannon entropy is given by:

$$H_I = -\sum_{i-1}^{c} \frac{s_i}{k} \log \frac{s_i}{k} \tag{1}$$

If the number of instances is exactly the same for all classes, Equation 1 reduces to $\log c$, and turns out that H_I makes sense only for $c > 1$. Therefore, if H_I is divided by $\log c$, it becomes normalized in the range [0..1], and we can obtain a measure of instances balance in the dataset, as follows:

$$B_I = \frac{H_I}{\log c} \tag{2}$$

Similar reasoning can be done considering the number of pixels of all samples in a class, so that we can obtain the pixels balance measure for the dataset, B_P. Consequently, to meet the above-mentioned definition of the (hypothetical) ideal dataset, $B_I = B_P = 1$. Since both, B_I and B_P, should be maximized, one could interpret them in the Cartesian plane, such that the farther from the origin, the better. Therefore, the Instances-Pixels Balance Index is defined as:

$$IPBI = \sqrt{B_I^2 + B_P^2} \tag{3}$$

B. THE PROPOSED METHOD

An overview of the proposed approach is shown in Figure 3, and it includes three steps: (1) Pre-processing, (2) Segmentation using EPYNET, and (3) Post-processing.

1) PRE-PROCESSING

The first step consists in detecting, localizing, and cropping all people from a raw image. For this task, we used the Single Shot MultiBox Detector (SSD) [18], which uses Convolutional Networks to generate bounding boxes for each class instance, in this case, people, only. Then, each person cropped from the raw image was proportionally resized to 320×320 (padding with zeros was applied to make it square). These transformations were also applied to the image label to ensure compatibility with respect to the input image. Finally, we used the min-max normalization to normalize the input images in the range [0..1], so that they are guaranteed to be on the same scale.

2) EPYNET: EFFICIENT PYRAMIDAL NETWORK

In this work, the clothing segmentation task was formulated as a classification problem. Therefore, a classifier was trained to assign each pixel to a target label. For this task, we propose a new approach, called EPYNET, base on FPN [19] architecture with the EfficientNet [20] model as the backbone.

FPN has achieved promising results in the object detection task, especially for small objects, by computing feature maps at different spatial resolutions in a pyramidal representation. This representation also can be used for semantic segmentation tasks (see, for instance, [34]). The EfficientNet is a family of models, ranging from the light-weight EfficientNet-B0 to the large EfficientNet-B7. It uses an effective compound scaling method to obtain improved efficiency and performance. Recently, the architecture achieved the state-of-the-art in image classification using the ImageNet dataset [20].

EPYNET performs the clothing segmentation task by combining these two approaches, taking advantage of both, the features extracted from the EfficientNet, and the pyramid representation at different scales from FPN. To better detail our proposed approach, shown in Figure 3, we divided the framework into three main components: Bottom-up Pathway, Top-down Pathway, and Semantic Segmentation Head.

The Bottom-up Pathway is a deep CNN, more specifically, the EfficientNet-B4 model initialized with weights trained on the ImageNet dataset. It consists of seven main convolution blocks, called MBConv, based on an Inverted Residual Block, previously introduced in MobileNet [35].

The Top-down Pathway consists of four blocks connected with lateral connections from the Bottom-up pathway that builds high-level semantic feature maps at different scales, using the image pyramid principle suggested by [19]. For the lateral connections, we use the following output layers from the EfficientNet-B4: block6a_expand_activation, block4a_expand_activation, block3a_expand_activation, block2a_expand_ctivation. Each block performs an upsampling by a factor of 2 from higher pyramids level features and adds with features from the Bottom-up pathway via lateral connections to locate the features more precisely.

Then, in the Semantic Segmentation Head, two convolutions with kernel size 3×3 and 128 channels are applied sequentially in each output block, followed by an upsampling operation to compute the final feature maps. For each convolution, we used batch normalization and ReLU (Rectified Linear Unit) activation function. All the four resulting feature maps were concatenated, and the last convolution with batch normalization and ReLU activation function is applied, followed by an upsampling operation.

Finally, the whole model was jointly optimized in an end-to-end way with the weighted cross-entropy loss to handle the imbalance in the training data. The model output is a 19-channel softmax layer that assigns to each pixel a value that represents the probabilities of belonging to each class of the dataset.

3) POST-PROCESSING

The output of the model consists of a softmax probability map of the same size of the input image with the number of channels defined by the number of classes. The opening morphological operation is applied to each channel of the predicted output to reduce noise impacts and enhance the detected regions. Finally, the prediction is merged using the argmax operation on each depth-wise pixel to obtain the final mask.

C. EVALUATION

Four measures were used to evaluate the proposed method: Precision, Recall, F1-score, and Intersection over Union (IoU), which are computed by Equations (4), (5), (6) and (7), respectively. These evaluation measures are commonly used for segmentation tasks in general and, also, for clothing segmentation [36]. All these measures are reported in this work aiming at comparisons with future approaches. It is worth to notice that Accuracy is not an appropriate performance measure for unbalanced datasets, as pointed out by [37]. Anyhow, Accuracy will be shown here only for the sake of comparing our approach with previously published works.

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{4}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{5}$$

$$\text{F1-Score}_i = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{6}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{7}$$

where TP, FP, FN indicate, respectively, True Positives, False Positives and False Negatives for a fiven classification i.

Precision is a metric that indicates the proportion of predictions that are true positives. Recall is a metric of completeness and specifies the proportion of positives that are detected. F1-score combines Precision and Recall and is the harmonic mean of these two metrics. IoU refers to the intersection of the ground truth and the predicted segmentation divided by the union of the ground truth and the predicted segmentation.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This Section presents experimental results obtained by the segmentation using the methods previously described in Section III. Firstly we describe the implementation details, the data augmentation techniques used during the training step, and the quantitative and qualitative results of the EPYNET on the UTFPR-SBD3 dataset. Then, a dataset comparison using the proposed IPBI measure is presented. Finally, the proposed approach is compared with other state-of-the-art approaches on CFPD, Fashionista, and Refined Fashionista datasets.

A. IMPLEMENTATION DETAILS

Experiments were performed on a workstation with Intel core i7-8700 processor, 32GBytes RAM, and a Nvidia Titan-Xp GPU. The Tensorflow and Keras library were used to train and test the proposed model.

We trained the EPYNET model with the UTFPR-SBD3 dataset. We used the RMSProp optimizer with default parameters and a learning rate of 0.001. During the training process, the learning rate was reduced by a factor of 0.1, whenever the evaluation metric did not improve for 5 epochs. A predefined number of 100 epochs was defined. However, training was stopped if the evaluation measure stagnated for 10 consecutive epochs.

The generalization performance of the trained model was accessed by means of the 10-fold cross-validation procedure, as usual in the literature [22], [29]. To ensure the same class distribution in each generated subset of the cross-validation, we used the stratified sampling method, proposed by [38]. Therefore, this procedure guarantees a less optimistic generalization estimate of the model.

By the end of each fold, the best model was used to predict the samples in the test set, and the performance was computed using the measures presented in Section III-C.

It is well-known that DL methods require massive volumes of data for training, and data augmentation techniques are frequently used to overcome the lack of large and high-quality datasets [39]. The basic idea is to make slight random changes in the input images to create more variety in the training data. This procedure is known to take more robustness to the trained models since they increase their generalization capability on unseen images (test dataset). Among the many methods that can be applied for image data augmentation [40], we used: flip, rotation, and random crop methods. There are two strategies for data augmentation, offline or online. In the first approach, the data augmentation methods are applied to the original training dataset to create a much larger dataset and, then, the augmented dataset is used for further training of the model. In the other approach, the data augmentation methods are randomly applied each time an image is presented to the model in the training step. We use the online approach as it requires less storage for the images, at the expense of some extra processing.

TABLE 3. Per-class segmentation performance obtained by the proposed model over the UTFPR-SBD3 dataset, with and without (w/o) including the Background class.

#	Class	Precision	Recall	F1-Score
1	Background	0.967	0.973	0.970
2	Bag	0.825	0.806	0.815
3	Belt	0.738	0.428	0.542
4	Coat	0.806	0.809	0.808
5	Dress	0.809	0.811	0.810
6	Eyewear	0.824	0.638	0.720
7	Footwear	0.836	0.827	0.832
8	Hair	0.858	0.890	0.874
9	Headwear	0.862	0.633	0.730
10	Neckwear	0.685	0.586	0.632
11	Pants	0.918	0.904	0.911
12	Romper/Jumpsuit	0.786	0.816	0.800
13	Shirt	0.785	0.793	0.789
14	Shorts	0.882	0.802	0.841
15	Skin	0.905	0.920	0.912
16	Skirt	0.820	0.848	0.834
17	Socks	0.762	0.648	0.700
18	Stocking	0.700	0.815	0.753
19	Sweater	0.673	0.566	0.615
	Average	0.813	0.764	0.783
	Average (w/o)	0.804	0.752	0.773

B. QUANTITATIVE RESULTS

The quantitative evaluation, presented in Table 3, shows the Precision, Recall, and F1-Score obtained with the proposed approach on UTFPR-SBD3 dataset. If including the Background as a class, our approach obtained 81.3%, 76.4%, and 78.3%, respectively. On the other hand, without the Background, values decrease slightly, although still suggesting that the inclusion of the Background among the classes of a clothing segmentation problem may lead to distorted results.

Considering the F1-Score, the best results were found for the following classes: Background, Skin, Pants, and Hair. On the other hand, the poorest results were those for classes related to small items: Belt, Sweater, Neckwear, Socks, Eyewear, and Headwear. These items are those with the smallest number of pixels among all classes. Moreover, Neckwear and Headwear are accessories that cover different styles and shapes (e.g., Neckwear class includes bowties, neckties, and scarves).

We also evaluate the overall segmentation performance of the model by using the IoU measure, as shown in Figure 4. The average IoU was 65.6%. According to [41], predictions with intersection over union more than 50% are considered satisfactory. Therefore, for most classes, results indicate that the proposed approach is efficient for the clothing segmentation task.

For only three classes the results were not satisfactory: Belt, Sweate, and Neckwear. Although the class Belt is present in approximately 30% of the dataset regarding the number of instances, it is a small object and occupies less than 1% of the dataset considering the number of pixels. Also, Neckwear and Sweater are among those classes with the lowest occurrence in the dataset.

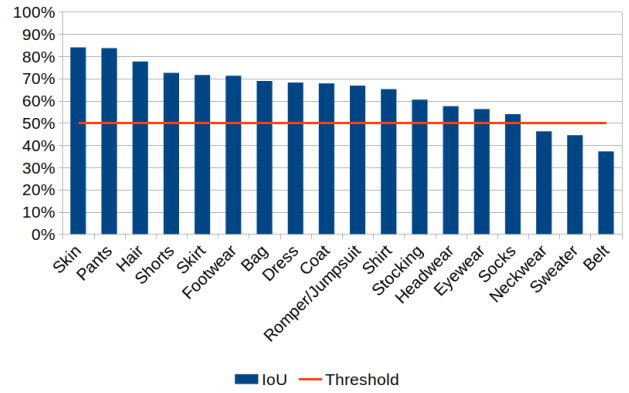

FIGURE 4. IOU scores for each class.

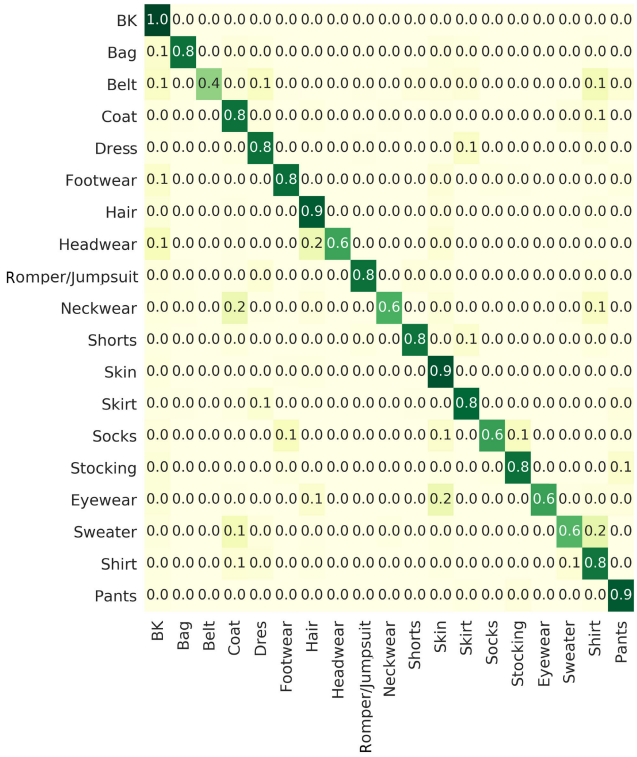

FIGURE 5. Confusion matrix computed on the UTFPR-SBD3 dataset from the sum of the confusion matrices generated in each fold of the cross-validation. Vertical axis: ground truth classes; Horizontal axis: predicted classes; cell values: percent of correct predictions.

The confusion matrix (Figure 5) shows the predicted and target classifications. For most classes, the highest values are found in the main diagonal, thus demonstrating the effectiveness of the proposed approach.

Even for classes of objects with small dimensions, such as Socks and Eyewear, the results were satisfactory, despite having some classification errors. For instance, Eyewear was occasionally incorrectly predicted as Skin or Hair, and Socks were wrongly predicted as Skin or Footwear. Although this is reasonable from the semantic point of view, this fact suggests that more data is required to improve the classification results. However, by increasing the number of instances of these classes, an inevitable increase in the number of pixels of other, more frequent classes (such as Skin and Hair) will occur. As a consequence, the class imbalance would increase instead of decreasing. We may also notice that many Sweaters were predicted as Shirt or Coat since those classes are quite similar to each other. Both types of misclassification highlight the complexity of the semantic segmentation of clothes, still a challenging problem.

C. QUALITATIVE RESULTS

In this Section, we present a visual evaluation of the outputs predicted by the proposed approach. Figure 7a) shows some sample images of the test set with both, the ground truth and the output provided by the EPYNET.

The trained model was able to satisfactorily segment different types of attributes, and confirm that our approach is robust enough for this task. Notwithstanding, in specific cases, poor segmentation results were also predicted by our model, as shown in Figure 7b). Notice that similar classes may cause misclassifications. For instance, Stocking was occasionally confused with Pants, Skirt was occasionally confused with Dress and Sweater may be confused as Shirt or Coat.

As mentioned in Section IV-B, for the Sweater class, segmentation results were not satisfactory. This class consists of a piece of clothing, made of knitted or crocheted material, that covers the top part of the body. It can be closed, also called a pullover, or opened, usually called a cardigan. Sweaters can have different shapes and styles and, due to such variability, this class was frequently predicted as Shirt or Coat. This may suggest that this category would be better divided into two or more classes (see Figure 6).

FIGURE 6. Segmentation results of the Sweater class.

Furthermore, the illumination variation, shadows, and partially occluded clothes may result in an incorrect segmentation. According to [42], the performance decreases

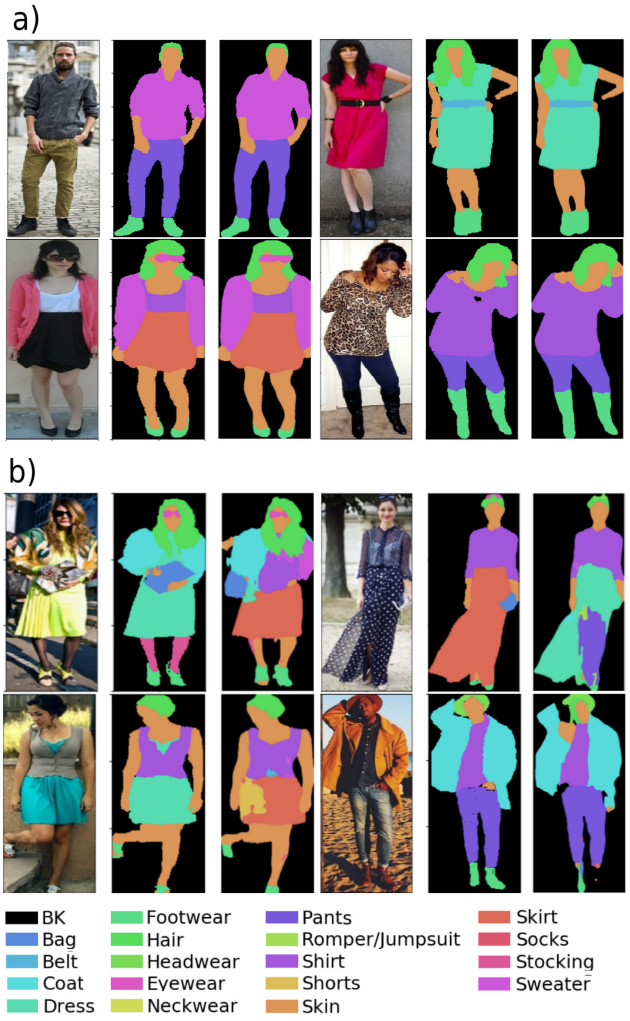

FIGURE 7. Segmentation results on the test set based on: a) highest and b) lowest IOU score. This Figure shows the input image, ground truth, and the predicted class from left to right, respectively.

significantly when more than 50% of a garment is occluded. For instance, Figure 7b) presents a dress that was predicted as a skirt, probably because it is partially occluded by a coat. Actually, such a problem is very difficult to address, even for humans.

D. COMPARISON OF DIFFERENT DATASETS

Based on the equations presented in Section III-A1, the Instances Balance (B_I), Pixels Balance (B_P) and IPBI were computed for all datasets used in this work. All metrics were computed with and without the Background class. This comparison is shown in Table 4.

In this Table, it is clear the effect of the Background class in the Pixels Balance (B_P) measure. Such class is the largest one (in pixels) in all datasets. Therefore, when present, it significantly decreases B_P (and, as consequence, *IPBI* too). No important changes are noticed in the Instances Balance (B_I) measure with or without the Background class.

TABLE 4. Measures of Instances Balance, Pixels Balance, and IPBI with (top lines) and without (botton lines) the Background class, for several clothing segmentation datasets.

	UTFPR-SBD3	CFPD	CCP	Fashionista	Ref.Fashionista
B_I	0.89	0.87	0.79	0.82	0.91
B_P	0.69	0.34	0.31	0.31	0.36
$IPBI$	0.79	0.66	0.60	0.62	0.69
$B_I(w/o)$	0.88	0.87	0.80	0.83	0.91
$B_P(w/o)$	0.84	0.86	0.76	0.78	0.86
$IPBI(w/o)$	0.86	0.86	0.78	0.81	0.89

TABLE 5. Comparison with the state-of-the-art approaches. For each dataset, the best results are highlighted.

Reference	Method	Acc	IoU
	CFPD		
[21]	Corful-Fashion		42.10
[44]	FCN-32s	90.34	47.65
[44]	FCN-16s	91.27	50.07
[44]	FCN-8s	91.58	51.28
[30]	Outfit filter	91.52	51.42
[43]	Deformable Network	93.06	53.51
[31]	SP-FEN	93.11	52.64
[32]	ResNetXt-FPN	93.82	54.39
(this work)	**EPYNET**	**93.83**	**59.52**
	Fashionista		
[22]	Paperdoll	84.68	
[45]	Clothelets CRF	84.88	
[44]	FCN-32s	85.94	29.61
[44]	FCN-16s	87.53	34.26
[44]	FCN-8s	87.51	33.97
[30]	Outfit filter	87.55	34.26
[43]	Deformable Network	89.21	37.12
(this work)	**EPYNET**	**89.80**	**41.09**
	Refined Fashionista		
[44]	FCN-32s	88.56	40.88
[44]	FCN-16s	89.74	43.96
[44]	FCN-8s	90.09	44.72
[30]	Outfit filter	91.50	46.40
[43]	Deformable Network	92.93	47.85
[31]	SP-FEN	91.83	47.07
[32]	ResNetXt-FPN	**93.62**	50.64
(this work)	**EPYNET**	92.06	**51.02**

Comparing the datasets according to *IPBI*, when including the Background class, notice that UTFPR-SBD3 is the best balanced-dataset, while the others are poorly balanced. Particularly, the imbalance of CCP and Fashionista is due to its large number of classes (more than 50), which many of them include small objects (such as gloves and rings), or have few instances. If excluding the Background class, UTFPR-SBD3 is similarly balanced as CFPD, quite close to the Refined Fashionista, which achieved the highest *IPBI*. However, recall that CFPD has many annotation errors caused by its superpixel-based annotation, as shown in Figure 1. The Refined Fashionista dataset, despite having the same number of images as the Fashionista, has improved the balance by around 10% when compared with the original dataset, thanks

to the fusion of classes with a small number of instances into larger ones.

E. COMPARISON WITH THE STATE-OF-THE-ART

The proposed approach was compared with the current state-of-the-art methods on the CFPD, Fashionista, and Refined Fashionista datasets. For a fair comparison, we use the same measures (Acc and IoU) reported by [30], available on GitHub.[3] Also, since the previous works did not exclude the Background class in the evaluation, the cropping step, described in Section III-B1, was not performed and the entire image was used as input during the training step.

The Fashionista dataset was divided into training and test set, as described in previous works [22], [30], [43], with 10% of training images left out for validation. On the other hand, the CFPD dataset was randomly divided into 78% for training, 2% for validation, and 20% for testing. Table 5 shows that, for all datasets, EPYNET achieved better results than the other methods in the literature.

Notice that our approach overpasses [32] by 9% of IoU in the CFPD dataset, and 1% in the Refined Fashionista dataset. In the Fashionista dataset, EPYNET also outperforms [44] by 11%. Despite the annotation issues reported in Section III-A, we achieved better results for accuracy in the CFPD and Fashionista datasets, and competitive results in the Refined Fashionista dataset.

V. CONCLUSION

Image segmentation has been one of the most challenging problems in computer vision that could be used to improve applications in many areas, including security and surveillance. Recently, soft biometrics traits, including types of clothes, have shown promising results in people's re-identification. However, it is still an open problem because of the wide variety of types, shapes, styles, and colors of clothes.

Although semantic segmentation using Deep Learning algorithms has achieved great success in many research fields, it is still difficult for computers to understand and describe a scene as humans naturally do.

As discussed in Section III-A1, the segmentation problem faced in this work is naturally unbalanced. The presence of classes with objects with a small number of pixels (e.g. belts, ties, or sunglasses), or that occurs in all images (e.g. skin and hair), makes it impossible to achieve a perfect balance. An imbalanced dataset, whether in terms of instances or pixels, can negatively influence the performance of segmentation methods.

This work has three contributions. First, motivated by the need for a large, high-quality dataset with pixel-level annotations, we created a new dataset, named UTFPR-SBD3. It was designated to overcome the annotation problems frequently found in other popular datasets, and to provide the best possible balance over classes at the instances and the pixels levels. Second, due to the difficulty in comparing datasets for clothing segmentation, a new measure of dataset imbalance was introduced: *IPBI*. With such a measure, it is possible to evaluate the influence of the background, classes with small items, or classes with a too high or too low number of instances. The third, and most important contribution, is EPYNET. This framework is based on the pyramidal architecture of FPN, and the EfficientNet model. It is aimed for clothing semantic segmentation, in the context of soft biometrics. We presented an extensive comparison of EPYNET with other approaches using several popular datasets, and it outperformed the state-of-art methods. Both, quantitative and qualitative results presented show the effectiveness of the EPYNET. Based on these results, we believe that the proposed approach can be potentially useful for many real-world applications related to soft biometrics, people surveillance, image description, clothes recommendation, people re-identification, and others.

Despite the good results achieved by EPYNET, we observed that other factors could influence the segmentation task, such as the environment illumination and the quality of the image. Besides, occlusions and similar classes in the dataset can degrade the predicted results. Future work will include improvements in the method to better handle illumination changes, and to enhance the discrimination between similar objects.

[3]https://github.com/pongsate1/fashion-parsing

ACKNOWLEDGMENT

The authors would like to thank to NVIDIA Corporation for the donation of the Titan-Xp GPU board used in this work.

REFERENCES

[1] M. Romero, M. Gutoski, L. T. Hattori, M. Ribeiro, and H. S. Lopes, "Soft biometrics classification in videos using transfer learning and bidirectional long short-term memory networks," *Learn. Nonlinear Models*, vol. 18, no. 1, pp. 47–59, Sep. 2020.

[2] A. Abdelwhab and S. Viriri, "A survey on soft biometrics for human identification," in *Machine Learning and Biometrics*, J. Yang, D. S. Park, S. Yoon, Y. Chen, and C. Zhang, Eds. Rijeka, Croatia: IntechOpen, 2018, ch. 3.

[3] X. Zhao, Y. Chen, E. Blasch, L. Zhang, and G. Chen, "Face recognition in low-resolution surveillance video streams," *Proc. SPIE*, vol. 11017, pp. 147–159, Jul. 2019.

[4] O. A. Arigbabu, S. M. S. Ahmad, W. A. W. Adnan, and S. Yussof, "Integration of multiple soft biometrics for human identification," *Pattern Recognit. Lett.*, vol. 68, pp. 278–287, Dec. 2015.

[5] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross, "Soft biometrics for surveillance: An overview," in *Handbook of Statistics*, vol. 31, C. Rao and V. Govindaraju, Eds. Amsterdam, The Netherlands: Elsevier, 2013, pp. 327–352.

[6] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 739–777, Jan. 2011.

[7] R. Vera-Rodriguez, P. Marin-Belinchon, E. Gonzalez-Sosa, P. Tome, and J. Ortega-Garcia, "Exploring automatic extraction of body-based soft biometrics," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–6.

[8] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018.

[9] S. Bashbaghi, E. Granger, R. Sabourin, and M. Parchami, *Deep Learning Architectures for Face Recognition in Video Surveillance*. Singapore: Springer, 2019, pp. 133–154.

[10] X. Di and V. M. Patel, "Deep learning for tattoo recognition," in *Deep Learning for Biometrics*. Cham, Switzerland: Springer, 2017, pp. 241–256.
[11] E. R. H. P. Isaac, S. Elias, S. Rajagopalan, and K. S. Easwarakumar, "Multiview gait-based gender classification through pose-based voting," *Pattern Recognit. Lett.*, vol. 126, pp. 41–50, Sep. 2019.
[12] H. A. Perlin and H. S. Lopes, "Extracting human attributes using a convolutional neural network approach," *Pattern Recognit. Lett.*, vol. 68, pp. 250–259, Dec. 2015.
[13] K. M. A. Raihan, M. Khaliluzzaman, and S. M. Rezvi, "Recognition of pedestrian clothing attributes from far view images using convolutional neural network," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–7.
[14] Z. Chen, S. Liu, Y. Zhai, J. Lin, X. Cao, and L. Yang, "Human parsing by weak structural label," *Multimedia Tools Appl.*, vol. 77, no. 15, pp. 19795–19809, Aug. 2018.
[15] C.-H. Yoo, Y.-G. Shin, S.-W. Kim, and S.-J. Ko, "Context-aware encoding for clothing parsing," *Electron. Lett.*, vol. 55, no. 12, pp. 692–693, Jun. 2019.
[16] W. Ji, X. Li, F. Wu, Z. Pan, and Y. Zhuang, "Human-centric clothing segmentation via deformable semantic locality-preserving network," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 25, 2019, doi: 10.1109/TCSVT.2019.2962216.
[17] E. S. Jaha and M. S. Nixon, "Soft biometrics for subject identification using clothing attributes," in *Proc. IEEE Int. Joint Conf. Biometrics*, Piscataway, NJ, USA, Sep. 2014, pp. 1–6.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[20] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 6105–6114.
[21] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2014.
[22] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3570–3577.
[23] A. D. S. Inácio, A. Brilhador, and H. S. Lopes, "Semantic segmentation of clothes in the context of soft biometrics using deep learning methods," in *Proc. 14th Brazilian Congr. Comput. Intell.*, Nov. 2020, pp. 1–7.
[24] A. Li, L. Liu, and S. Yan, *Person Re-Identification by Attribute-Assisted Clothes Appearance*. London, U.K.: Springer, 2014, pp. 119–138.
[25] M. Ribeiro, M. Gutoski, A. E. Lazzaretti, and H. S. Lopes, "One-class classification in images and videos using a convolutional autoencoder with compact embedding," *IEEE Access*, vol. 8, pp. 86520–86535, 2020.
[26] P. Hu, D. Peng, Y. Sang, and Y. Xiang, "Multi-view linear discriminant analysis network," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5352–5365, Nov. 2019.
[27] J.-S. Zhang, J. Cao, and B. Mao, "Application of deep learning and unmanned aerial vehicle technology in traffic flow monitoring," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2017, pp. 189–194.
[28] T. Hrkac, K. Brkic, and Z. Kalafatic, "Multi-class U-Net for segmentation of non-biometric identifiers," in *Proc. 19th Irish Mach. Vis. Image Process. Conf.*, 2017, pp. 131–138.
[29] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA, Jun. 2014, pp. 3182–3189.
[30] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," 2017, *arXiv:1703.01386*. [Online]. Available: https://arxiv.org/abs/1703.01386
[31] A. M. Ihsan, C. K. Loo, S. A. Naji, and M. Seera, "Superpixels features extractor network (SP-FEN) for clothing parsing enhancement," *Neural Process. Lett.*, vol. 51, pp. 1–19, Jan. 2020.
[32] J. Martinsson and O. Mogren, "Semantic segmentation of fashion images using feature pyramid networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3133–3136.
[33] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu, "ModaNet: A large-scale street fashion dataset with polygon annotations," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, 2018, pp. 1670–1678.
[34] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 272–275.
[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[36] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.
[37] G. E. A. P. A. Batista, A. C. P. L. F. Carvalho, and M. C. Monard, "Applying one-sided selection to unbalanced datasets," in *Proc. Mex. Int. Conf. Artif. Intell.* Berlin, Germany: Springer, 2000, pp. 315–325.
[38] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2011, pp. 145–158.
[39] N. Aquino, M. Gutoski, L. Hattori, and H. Lopes, "The effect of data augmentation on the performance of convolutional neural networks," in *Proc. 13th Brazilian Conf. Comput. Intell. (SBIC/ABRICOM)*, 2017, pp. 1–12.
[40] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 60, pp. 1–48, 2019.
[41] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[42] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5332–5340.
[43] W. Ji, X. Li, Y. Zhuang, O. E. F. Bourahla, Y. Ji, S. Li, and J. Cui, "Semantic locality-aware deformable network for clothing segmentation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 764–770.
[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[45] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "A high performance CRF model for clothes parsing," in *Proc. Asican Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2015, pp. 64–81.

ANDREI DE SOUZA INÁCIO received the B.Sc. and M.Sc. degrees in computer science from the Federal University of Santa Catarina (UFSC), in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Federal University of Technology – Paraná, Brazil. Since 2014, he has been a Lecturer with the Federal Institute of Santa Catarina (IFSC). He has professional experiences in information systems design, Web development, and IT project management. His research interests include, but not limited to computer vision, machine learning, and data mining.

HEITOR SILVÉRIO LOPES received the B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Technology – Paraná (UTFPR), Curitiba, in 1984 and 1990, respectively, and the Ph.D. degree from the Federal University of Santa Catarina, in 1996. Since 2003, he has been a Research Fellow with the Brazilian National Research Council, in the area of computer science. In 2014, he spent a sabbatical year at the Department of Electrical Engineering and Computer Science, The University of Tennesse, Knoxville, USA. He is currently a Tenured Full Professor with the Department of Electronics and the Graduate Program on Electrical Engineering and Applied Computer Science (CPGEI), UTFPR. He is also the Founder and the Current Head of the Bioinformatics and Computational Intelligence Laboratory (LABIC). His research interests include computer vision, deep learning, evolutionary computation, and data mining.