*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1

*Article*

# Human Attribute Recognition: A Comprehensive Survey

**Ehsan Yaghoubi [1,†,*], Farhad Khezeli[2], Diana Borza[3], SV Aruna Kumar[4], João Neves[5], and Hugo Proença[1,*]**

1 IT: Instituto de Telecomunicações, University of Beira Interior; ehsan.yaghoubi@ubi.pt; hugomcp@di.ubi.pt
2 Science and Research Branch, Islamic Azad University; farhad.khezeli@srbiau.ac.ir
3 Technical University of Cluj-Napoca; diana.borza@cs.utcluj.ro
4 University of Beira Interior; aruna.venkateshappa@ubi.pt
5 TomiWorld; joaoneves@tomiworld.com
* Correspondence: ehsan.yaghoubi@ubi.pt; Tel.: (+351-915840758); hugomcp@di.ubi.pt
† Current address: SOCIA Lab., Computer Science Faculty, University of Beira Interior, Portugal.

**Abstract:** Over the last decade, the field of Human Attribute Recognition (HAR) has dramatically changed, mainly due to the improvements brought by deep learning solutions. This survey reviews the progress obtained in HAR, considering the transition from the traditional hand-crafted to deep-learning approaches. The most relevant works on the field are analyzed concerning the advances proposed to address the HAR's typical challenges. Furthermore, we outline the applications and typical evaluation metrics used in the HAR context. Finally, we provide a comprehensive review of the publicly available datasets for the development and evaluation of novel HAR approaches.

**Keywords:** Human Attribute Recognition; Imbalanced Learning; Pedestrian Recognition; Privacy Concerns; Clothing Attributes; Soft Biometrics; Appearance-Based Learning; Deep Learning.

## 1. Introduction

Over the recent years, the increasing amounts of multimedia data collected from the internet or by cameras deployed in public/private environments has been raising the requirements for solutions able to automatically analyse human appearance, features and behavior. Hence, human attribute analysis (HAR) has been attracting increasing attentions in the computer vision/pattern recognition domains, mainly due to its potential usability for a wide range of applications (e.g., crowd analysis, person search, detection, tracking, and re-identification). Human attribute analysis aims at describing and understanding the subjects' traits (such as their hair color, clothing style, age, etc.) either from full-body or facial data. Generally, there are four main sub-categories in the area of human attribute analysis:

- **Facial attribute analysis**. Facial attribute analysis aims at is at estimating the facial attributes or manipulating the desired attributes. The former is usually carried out by extracting a comprehensive feature representation of the face image, followed by a classifier to predict the face attributes. On the other hand, in manipulation works, face images are modified (e.g., glasses are removed or added) using generative models.
- **Full-body attribute recognition**. Full-body attribute recognition regards the task of inferring the soft-biometric labels of the subject, including clothing style, head-region attributes, recurring actions (talking to the phone) and role (cleaning lady, policeman), regardless of the location or body position (eating in a restaurant).

 doi:10.20944/preprints202007.0055.v1



- **Pedestrian attribute recognition**. As an emerging research sub-field of HAR, it focuses on the full-body human data that have been exclusively collected from video surveillance cameras or panels, where persons are captured while walking, standing, or running.
- **Clothing attribute analysis**. Another sub-field of human attribute analysis that is exclusively focused on clothing style and type. It comprises several sub-categories such as in-shop retrieval, costumer-to-shop retrieval, fashion landmark detection, fashion analysis, and cloth attribute recognition, each of which requires specific solutions to handle the challenges in the field. Among these sub-categories, cloth attribute recognition is similar to pedestrian and full-body attribute recognition and studies the clothing types (e.g., texture, category, shape, style).

Regarding the previously published surveys that addressed similar topics, we particularly mention Zheng et al. [1], where the facial attribute manipulation and estimation methods have been reviewed. However, to date, there is no solid survey on the recent advances in other sub-categories of human attribute analysis. As the essence of full-body, pedestrian, and cloth attribute recognition methods are similar to each other; in this paper, we cover all of them with a particular focus on the pedestrian attribute recognition methods. Meanwhile, [2] is the only work similar to our survey that is about pedestrian attribute recognition. Several points distinguish our work from [2]:

- The recent literature on HAR has been mostly focused in addressing some particular challenges of this problem (such as class imbalance, attribute localization, etc.) rather devising a general HAR system. Therefore, instead of providing a methodological categorization of the literature as in [2], our survey proposes a challenge-based taxonomy, discussing the state-of-the-art solutions and the rationale behind them;
- Contrary to [2], we analyze the motivation of each work and the intuitive reason for its superior performance;
- The datasets main features, statistics and kinds of annotation are compared and discussed in detail;
- Beside the motivations, we discuss HAR applications, divided into three main categories: security, commercial, and related research directions.

The typical pipeline of the HAR systems is given in Fig. 1, which indicates the requirement of a dataset preparation prior to designing a model. As shown in Fig. 1, preparing a dataset for this poblem typically comprises four steps:

1. Capturing raw data, which can be accomplished using mobile cameras (e.g., drone) or stationary cameras (e.g., CCTV). Also, the raw data might even be collected from images/videos publicly available (e.g., *Youtube*, or similar sources).
2. In most supervised training approaches, HAR models consider one person at a time (instead of analyzing a full-frame with multiple persons). Therefore, detecting the bounding boxes of each subject is essential and can be done by state-of-the-art object detection solutions (i.e., MaskRCNN, YOLO, SSD, etc.)
3. If the raw data is in video format, spatio-temporal information should be kept. in such cases, the accurate tracking of each object (subject) in the scene can significantly ease the annotation process.
4. Finally, in order to label the data with semantic attributes, all the bounding boxes of each individual are displaced to human annotators. based on human perception, the desired labels (e.g., 'gender' or 'age') are then associated to each instance of the dataset.

Regarding the data-type and available annotations, there are many possibilities for designing HAR models. Early researches were based on crafted feature extractors. Typically, the linear support vector machine (SVM) was used with different descriptors (such as ensemble of localized features, local binary patterns, color histograms, histogram of oriented gradients) to estimate the human attributes. However, as the correlation between human attributes were ignored in traditional methods, one single model was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



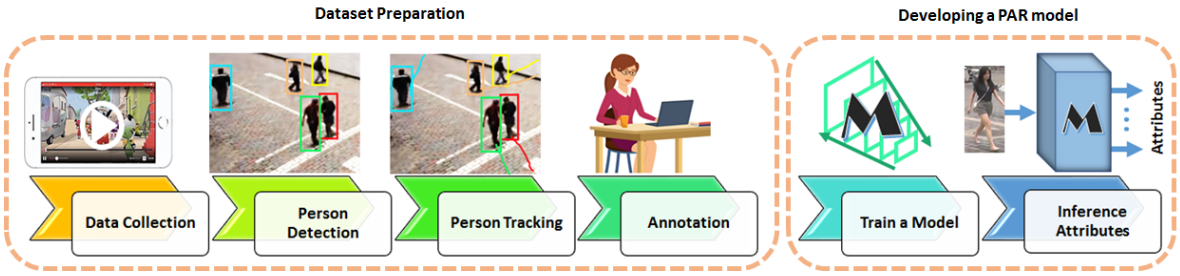


**Figure 1.** Typical pipeline to develop a HAR model.

not suitable for estimating several attributes. For instance, descriptors suitable for gender recognition could not be effective enough to recognize the hairstyle. Therefore, conventional methods focused on constructing independent feature extractors for each attribute. After the advent of convolutional neural networks(CNNs) and using it as a holistic feature extractor, a growing number of methods focused on models that can estimate multiple attributes at once. Earlier deep-based methods used shallow networks (e.g., 8-layer AlexNet [3]), while later models moved towards deeper architectures (e.g., residual network(ResNet)) [4]. Several major advantages of deep learning approaches moved the main research trend towards the deep neural network methods. First, CNNs are end-to-end (i.e., both the feature extraction and classification layers are trained simultaneously). Second, the deep neural networks' high generalization ability has provided the possibility of transferring the knowledge of other similar fields to scenarios with limited data. As an example, applying the weights of a model that has been trained on a large dataset (e.g., ImageNet) not only has shown positive effects on the accuracy of the model but also has decreased the convergence time and over-fitting problem [5], [6], [7]. Thirdly, CNNs could be designed to handle multiple tasks and labels in a unified model [8], [9].

Despite the large advances, HAR is still an open problem, mainly due to the high-variability in appearance that intra-class examples possess. Nevertheless, the following factors have been identified as the basis for the development of robust HAR systems:

- learn in an end-to-end manner and yield multiple attributes at once;
- extract a discriminative and comprehensive feature representation from the input data;
- leverage the intrinsic correlations between attributes;
- consider the location of each attribute in a weakly supervised manner;
- are robust to primary challenges such as low-resolution data, pose variation, occlusion, illumination variation, and cluttered background;
- handle the classes imbalance;
- manage the limited-data problem effectively.

### *1.1. Motivation And Applications*

Human attribute recognition methods extract semantic features that describe human-understandable characteristics of the individuals in a scene, either from images or video sequences, ranging from demographic information (gender, age, race/ethnicity), appearance attributes (body weight, face shape, hairstyle and color etc.), emotional state, to the motivation and attention of people (head pose, gaze direction). As they provide vital information about humans, such systems have already been integrated into numerous real-world applications, and are entwined with many technologies across the globe.

Indisputably, HAR is one of the most important steps in any visual surveillance system. Biometric identifiers are extracted to identify and distinguish between the individuals. Based on the biometric traits, humans are uniquely identified, either based on their facial appearance [10–12], iris patterns [13] or on behavioral traits (gait) [14,15]. With the increase of surveillance cameras worldwide, the research focus has shifted from (hard-)biometric (iris recognition, palm-print) to soft biometric identifiers. The latter describe human characteristics, taxonomized into a humanly understandable manner, but are

 doi:10.20944/preprints202007.0055.v1



not sufficient to uniquely differentiate between individuals. Instead, they are descriptors used by humans to categorize their peers into several classes.

On a top level, HAR applications can be divided into three main categories: *security and safety*, *research* directions, and *commercial applications*.

Yielding high-level semantic information, HAR could provide auxiliary information for different computer vision tasks, such as person re-identification ([16], [17]), human action recognition [18], scene understanding, advanced driving assistance systems, and event detection ([19]).

Another fertile field where HAR could be applied is in human drone surveillance. Drones or Unmanned Aerial Vehicles (UAV), although initially designed for military applications, are rapidly extending to various other application domains, due to their reduced size, swiftness, and ability to navigate through remote and dangerous environments. Researchers in multiple fields have started to use UAVs drones in their research work, and, as a result, the Scopus database has shown an increase in the papers related to UAVs, from 11 ($4.7 \times 10^6$ of total papers) papers published in 2009 to 851 ($270.0 \times 10^6$ of total articles) published in 2018 [20]. In terms of human surveillance, drones have been successfully used in various scenarios, ranging from rescue operations and victim identification, people counting and crowd detection, to police activities. All these applications require information about human attributes.

Nowadays, researchers in universities and major car industries work together to design and build the self-driving cars of the future. HAR methods have important implications in such systems as well. Although numerous papers addressed the problem of pedestrian detection, pedestrian attribute recognition is one of the keys to future improvements. Cues about the pedestrians' body and head orientation provide insights about their intent, and thus avoiding collisions. The pedestrians' age is another aspect that should be analyzed by advanced driving assistance systems to decrease vehicle speed when children are on the sidewalk. Finally, other works suggest that even pedestrians' accessories could be used to avoid collisions: starting from the statistical evidence that collisions between pedestrians and vehicles are more frequent on rainy days, in [21] authors suggest that detecting whether a pedestrian has on open umbrella could reduce traffic incidences.

As mentioned above, the applications of biometric cues are not limited to surveillance systems. Such traits have necessary implications also in commercial applications (logins, medical records management) and government applications (ID cards, border, and passport control) [22]. Also, a recent trend is to have advertisement displays in malls and stores equipped with cameras and HAR systems to extract socio-demographic attributes of the audience and present appropriate and targeted ads based on the audience's gender, generation or age.

Of course, this application list is not exhaustive, and numerous other practical uses of HAR can be envisioned, as this task has implications in all fields interested in and requiring (detailed) human description.

## 2. Discussion of Sources

As depicted in Fig 2, we identified five major challenges frequently addressed by the literature of HAR: data class imbalance, limited learning data, attribute relation, part occlusion, and localization. Among these challenges, considering the attribute correlation and extracting fine-grained features from local regions of the given data have attracted the most attention. Although data class imbalance has not been the main contribution of many works, it is a major concern in HAR and is often handled by applying weighted loss functions. To deal with limited data challenges, scholars frequently apply the existing holistic transfer learning and augmentation techniques in computer vision and pattern recognition. In this section, we discuss the significant contributions of the literature works in alleviating the above-mentioned challenges in HAR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



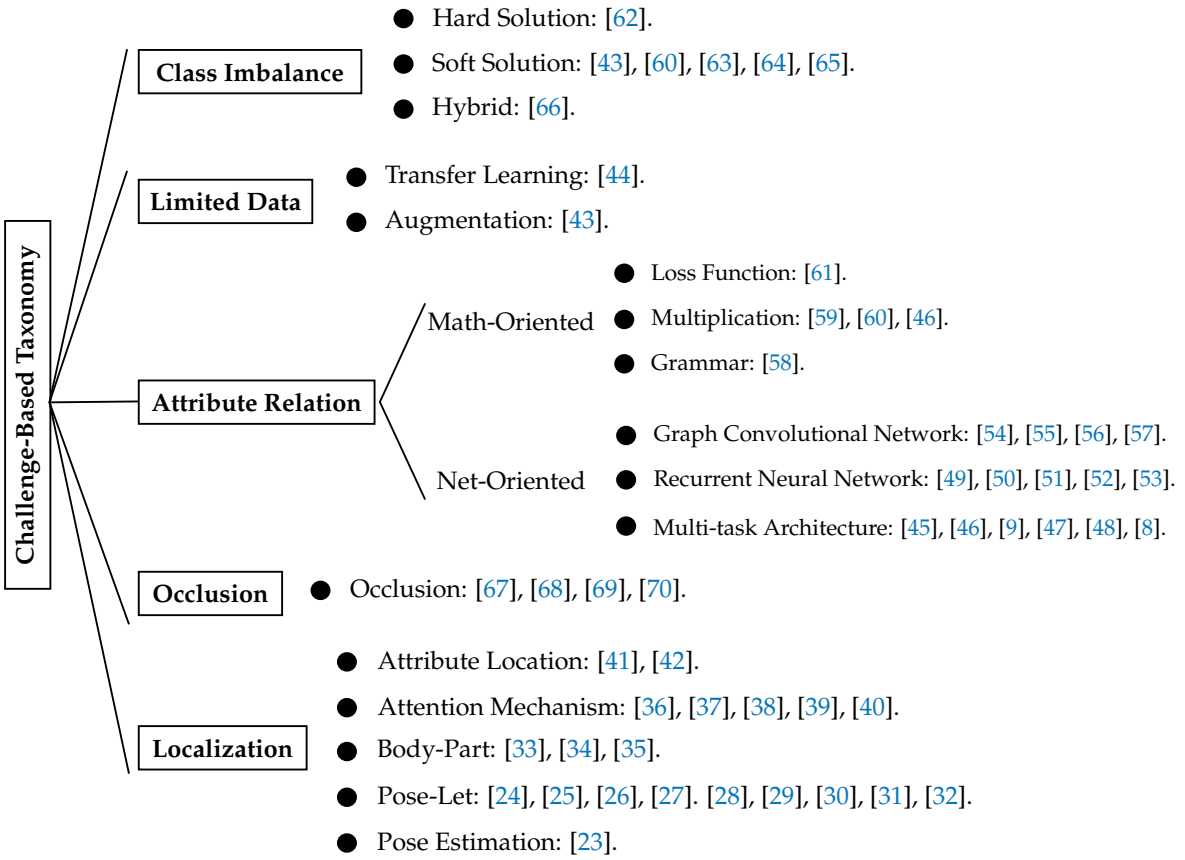


**Figure 2.** Proposed Taxonomy for main challenges in HAR.

### *2.1. Localization Methods*

Analyzing human full-body images only yields the global features; therefore, to extract distinct features from each identity, analyzing local regions of the image becomes important [71]. To capture the human fine-grained features, typical methods divide the person's image into several strides or patches and aggregate all the decisions on parts to yield the final decision. Further, decomposition of human-body and comparing it with others is intuitively similar to localizing the semantic body-parts and then describing them. Although the extracted localized features create a detailed feature representation of the image, dividing the image to several pieces has several drawbacks:

- the expressiveness of the data is lost (e.g., when processing a jacket only by several parts, some global features that encode the jacket's shape and structure are ignored).
- as the person detector cannot always provide aligned and accurate bounding boxes, rigid partitioning methods are prone to error in body-part captioning, mainly when the input data includes a wide background.
- different from gender and age, most human attributes (such as glasses, hat, scarf, shoes, etc.) belong to small regions of the image; therefore, analyzing other parts of the image is unnecessary and may add irrelevant features to the final feature representation of the image.
- some attributes are view-dependent and highly changeable due to human body-pose, and ignoring them reduces the model performance; for example, glasses recognition in the side-view images is more laborious than front-view, while it may be impossible in back-view images.

There are several possibilities to address these issues, and all of them try to guide the learning process using additional information. For instance, some works use novel model structures [31] to capture the relationships and correlations between the parts of the image, while others try to use prior body-pose coordinates [23] (or develop a view-detector in the main structure [72]) to learn the

 doi:10.20944/preprints202007.0055.v1



view-specific attributes. Some methods develop attention modules to find the relevant body parts, while some approaches extract various pose-lets[73] of the image by slicing-window detectors. Using the semantic attributes as a constraint for extracting the relevant regions is another solution to look for localized attributes [42]. Moreover, developing accurate body-part detectors and introducing datasets with part annotations are two strategies that can help the localization methods. In the following, we review the literature of HAR methods based on parts-localization and discuss their effectiveness in improving the baseline performance.

#### 2.1.1. Pose Estimation-Based Methods

Considering the effect of the body-pose variation of the feature representation, [72] proposes to learn multiple attribute classifiers so that each of them is suitable for a specific body-pose. Therefore, authors use the Inception architecture [74] as the backbone feature extractor, followed by three branches to capture the specific features of the front, back, and side views of the individuals. Simultaneously, a view-sensitive module analyzes the extracted features from the backbone to refine each branch's scores. The final results are the concatenation of all the scores. Ablation studies on the PETA dataset show that a plain Inception model achieves an 84.4 $F1$-score, while for the model with a pose-sensitive module, this metric increases to 85.5.

[23] is another research that takes advantage of pose estimation for improving the performance of pedestrian attribute recognition. In this work, Li et al. suggested a two-stream model whose results are fused, allowing the model to benefit from both regular global and pose-sensitive features. Given an input image, the first stream extracts the regular global features. The pose-sensitive branch comprises three steps: 1) coarse pose estimator (body-joint coordinates predictor) by applying the approach proposed in [75], 2) region localization that uses the body-pose information to spatially transform the desired region, originally proposed in [76], 3) fusion layer that concatenates the features of each region. In the first step, pose coordinates are extracted to be shared with the second module, in which body parts are localized by using spatial transformer networks [77]. A specific classifier is then trained for each region. Finally, the extracted features from both streams are concatenated to return a comprehensive feature representation of the given input data.

#### 2.1.2. Pose-Let-Based Methods

The main idea of pose-let based methods is to provide a bag-of-features from the input data using different patching technique. As earlier methods lacked accurate body part detectors, overlapping patches of the input images were used to extract local features.

[27] is one of the first techniques in this group that uses Spatial Pyramid Representation (SPR) [78] to divide the images into grids. Unlike a standard bag-of-features method that extracts the features from a uniform patching distribution, they suggest a recursive splitting technique, in which each grid has a parameter that is jointly learned with the weight vector. Intuitively, the spatial grids are varying for each class, which leads to better feature extraction.

In [24], hundreds of pose-lets are detected from the input data; a classifier is trained for each pose-let and semantic attribute. Then, another classifier aggregates the body-part information, with emphasis on the pose-lets taken from usual viewpoints that have discriminative features. A third classifier is then used to consider the relationship between the attributes. This way, by using the obtained feature representation, the body pose and viewpoint are implicitly decomposed.

Noticing the importance of accurate body-part detection when dealing with clothing appearance variations, [25] proposes to learn a comprehensive dictionary that considers various appearance part types (e.g., representing the lower-body in different appearances from bare legs to long skirts). To this end, all the input images are divided into static overlapping cells, each of which is represented by a feature descriptor. Then, as a result of feature clustering into $K$ clusters, they represent $k$ types of appearance parts.

 doi:10.20944/preprints202007.0055.v1



In [28], the authors targeted the human attributes and action recognition from still images. To this end, supposing that the available human bounding boxes are located in the center of the image, the model learns the scale and positions of a series of image partitions. Later, the model predicts the labels based on the reconstructed image from the learned partitions.

To address the large variation in articulation, angle, and body-pose [30] proposes a CNN-based features extractor, in which each pose-let is fed to an independent CNN. Then, a linear SVM classifier learns to distinguish the human attributes based on the aggregation between the full-body and pose-let features.

[31] and [32] showed that not only CNNs can yield a high-quality feature representation from the input, but also they are better at classification than SVM classifiers. In this context, Zhu et al. propose to predict multiple attributes at-once, by implicit regard to the attribute dependencies. Therefore, the authors divide the image into 15 static patches and analyze each one with a separate CNN. To consider the relationship between attributes and patches, they connect the output of some specific CNNs to the relevant static patches. For example, the upper splits of the images are connected to the head and shoulder's attributes.

[29] claims that in previous pose-let works, the location information of the attributes is ignored. For example, to recognize whether a person wears a hat or not, knowing that this feature is related to the upper regions of the image can guide the model to extract more relevant features. To implement this idea, the authors used an Inception [74] structure, in which the features of three different levels (low, middle, and high levels) are fed to three identical modules. These modules extract different patches from the whole and part of the input feature maps. The aggregation of the three branches constructs the final feature representation. By following this architectural design, the model implicitly learns the regions related to each attribute in a weakly supervised method. Surprisingly, the baseline (the same implementation without the proposed module) achieves better results on the PETA dataset (84.9 vs. 83.4 of $F1$), while on RAP dataset, the results of the model equipped with their proposed module ($F168.6$) is better with a margin of 2.

[26] receives the full frames and uses the scene features (i.e., hierarchical contexts) to help the model learn the attributes of the targeted person. For example, in a sports scene, it is expected that people have sporty style clothing. Using Fast R-CNN [79], the bounding box of each individual is detected, and several pose-let are extracted. After feeding the input frame and its Gaussian pyramids into several convolutional layers, four fully connected branches are added to the top of the network to yield four scores (from human bounding box, pose-lets, nearest neighbors of the selected parts, and full-frame) for a final concatenation.

#### 2.1.3. Part-Based Methods

Extracting discriminative fine-grained features often requires first to localize patches of the relevant regions in the input data. Unlike pose-let-based methods that detect the patches from the entire image, part-based methods aim to learn based on accurate body parts (i.e., head, torso, arms, and legs). Optimal part-based models are 1) pose sensitive (i.e., for similar poses, shows strong activations), 2) extendable to all samples, 3) discriminative on extracting features. CNNs can handle all these factors to some extend, and [33] empirical experiments confirm that for deeper networks, accurate body-parts are less significant.

As one of the first part-based works, inspired by a part detector (i.e., deformable part model [80], which captures viewpoint and pose variations), Zhang et al. [35] propose two descriptors that learn based on the part annotations. Their main objective is to localize the semantic parts and construct a normalized pose. To this end, the first descriptor is fed by correlated body parts, while for the second descriptor, the input body splits have no semantic correlation. Intuitively, the first descriptor is based on the inherent semantics of the input image, and the second descriptor learns the cross-component correspondences between the body parts.

 doi:10.20944/preprints202007.0055.v1



Later, in this context [33] proposes a model composed of a CNN-based body-part detector, including an SVM classifier (trained on the full-body and body parts, i.e., head, torso, and legs) to predict the human attributes and action. Given an input image, a Gaussian pyramid is constructed, each level is fed to several convolutional layers to produce pyramids of feature maps. The convolution of each feature-level with each body-part produces scores correspond to that body-part. Therefore, the final output is a pyramid of part model scores suitable for learning an SVM classifier. The experiments indicate that using body-part analysis and making the network deeper improve the results.

As earlier part-based methods used separate feature extractors and classifiers, the parts could not be optimized for recognizing the semantic attributes. Moreover, the detectors, at that time, were inaccurate in detection. Therefore, [34] proposed an end-to-end model, in which the body partitions are generated based on the skeleton information. As authors augment a large skeleton estimation dataset (MPII[81]) for human skeleton information (which is less prone to error for annotation in comparison with bounding box annotations for body parts), their body detector is more accurate in detecting the relevant partitions, leading to better performance.

To encode both global and fine-grained features and implicitly relate them to the specific attributes, [82] proposes to add several branches on top of a ResNet50 network, such that each branch explores particular regions of the input data and learns an exclusive classifier. Meanwhile, before the classifier stage, all branches share a layer, which passes the 6 static regions of features to the attribute classifiers. For example, the head attribute classifier is fed only with the two upper strips of the feature maps. Experimental results on the Market-1501 dataset [17] show that applying a layer that feeds regional features to the related classifiers can improve the $mA$ from 85.0 to 86.2. Further, repeating the experiments while adding a branch to the architecture of the model for predicting the person ID (as an extra-label) improves the $mA$ result from 84.9 to 86.1. These experiments show that simultaneous ID prediction without any purpose could slightly diminish the accuracy.

#### 2.1.4. Attention Based Methods

By focusing on the most relevant regions of the input data, human beings recognize the objects and their attributes without the background's interference. For example, when recognizing the head-accessories attributes of an individual, special attention is given to the facial region. Therefore, many HAR methods have attempted to implement an attention module to be inserted at multiple levels of CNN. Attention heat maps are colorful localization score maps that make the model interpretable and are usually faded over the original image to show the model's ability to focus on the relevant regions.

In order to eliminate the need for body-part detection and prior correspondence among the patches, [36] proposed to refine the Class Activation Map network [83], in which the relevant regions of the image to each attribute are highlighted. The model comprises a CNN feature extraction backbone with several branches on its top, which yield the scores for all the attributes and their regional heat maps. The fitness of the attention heat maps is measured using an exponential loss function, while the score of the attributes is derived from a classification loss function. The evaluation of the model is performed using two different convolutional backbones (i.e., VGG16 [84] and AlexNet [3]), and the result for the deeper network (VGG16) is better than the other one.

To extract more distinctive global and local features, Liu et al. [37] propose an attention module that fuses several feature layers of the relevant regions and yields attention maps. To take full advantage of the attention mechanism, they apply the attention module to different model levels. Obtaining the attentive feature maps from various layers of the network means that the model has captured multiple levels of the input sample's visual patterns so that the attention maps from higher blocks can cover more extensive regions, and the lower blocks focus on smaller regions of the input data.

Considering the problem of cloth classification and landmark detection, [39] proposes an attentive fashion grammar network, in which both the symmetry of the cloths and effect of body motion is captured. To enhance the clothing classification, authors suggest to 1) develop supervised attention

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



using the ground truth landmarks to learn the functional parts of the clothes and 2) use a bottom-up, top-down network [85], in which a successive down and up-sampling are performed on the attention maps to learn the global attention. The evaluation results of their model for clothing attribute prediction improved the counterpart methods by a large margin (30% to 60% top-5 accuracy on the DeepFashoin-C dataset [86]).

With a view to select the discriminative regions of the input data, [38] proposes a model considering three aspects: 1) Using the parsing technique [87], they split features of each body-part and help the model learns the location-oriented features by pixel-to-pixel supervision. 2) Multiple attention maps are assigned to each label due to empowering the features from the relevant regions to that label and suppressing the other features. Different from the previous step, the supervision in this module is performed on the image-level. 3) Another module learns the relevant regions for all the attributes and learns from a global perspective. The quantitative results on several datasets show that the full version of the model improves the plain model's performance slightly (e.g., for the RAP dataset, the F1 metric improves from 79.15 to 79.98).

[40] is another research that has focused on localizing the human attributes engaging multi-level attention mechanisms in full-frame images. First, supervised coarse learning is performed on the target person, in which the extracted features of each residual block is multiplied by the ground truth mask. Then, inspired by [88], to further boost the attribute-based localization, an attention module uses the labels to refine the aggregated features of multiple levels of the model.

To alleviate the complex background and occlusion challenges in HAR, [89] introduces a coarse attention layer that uses the multiplication between the output of the CNN backbone and ground truth human masks. Further, to guide the model to consider the semantic relationships among the attributes, authors use a multi-task architecture with a weighted loss function. This way, the CNN learns to find the relevant regions to the attributes in the foreground regions. Their ablation studies show that considering the correlation between attributes (multi-task learning) is more effective than coarse attention on the foreground region, although both improve the model performance.

#### 2.1.5. Attribute Based Methods

Noticing the effectiveness of the additional information (e.g., pose, body-part and viewpoint) in the global feature representation, [41] introduces a method that improves the localization ability of the model by locating the attributes' regions in the images. The model comprises two branches, one of them extracts the global features and provides the CAMs [90] (attention heat-maps), and the other one uses [91] to produce some regions of interests (RoI) for extracting the local features. To localize each attribute, the authors consider regions with high overlap between the CAMs and RoIs as the attribute location. Finally, the local and global features are aggregated using an element-wise sum. Their ablation studies on the RAP dataset show that for the model without localization $F1$ metric is about 77%, while the full-version model improves the results to about 80%.

As a weakly supervised method, [42] aims to learn the regions in the input data related to the specific attributes. Thereby, the input image is fed into a BN-Inception model [92], and the features from three levels of the model (low, mid, and high) are concatenated together to be ready for three separate localization process. The localization module is built from a SE-Net [93] (that considers the channel relationships) proceeded with a Spatial Transformer Network (STN) (that performs conditional transformations on the feature maps) [77]. The training is weakly supervised because instead of using the ground truth coordinates of the attribute region, the STN is treated as a differentiable RoI pooling layer that is learned without box annotations. The $F1$ metric on the RAP dataset for BN-Inception plain model is around 78.2 while this number fro the full version of the model is 80.2.

Considering that both the local and global features are important for making a prediction, most of the literature's localization-based methods have introduced modular techniques. Therefore, the proposed module could be used in multiple levels of the model (from the first convolutional layers

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



to the final classification layers) to capture both the low-level and high-level features. Intuitively, the implicit location of each attribute is learned in a weakly supervised manner.

### 2.2. Limited Data

Although deep neural networks are powerful in the attribute recognition task, an insufficient amount of data causes an early overfitting problem and hinders them from extracting a generalized feature representation from the input data. Meanwhile, the deeper the networks are, the more data are required to learn a wide range of layer weight parameters. The primary solutions for solving the problem fo limited data are synthesizing artificial samples or augmenting the original data. One of the popular approaches for increasing the size of the dataset is to use generative models (i.e., Generative Adversarial Network(GAN)[94], Variational Auto-Encoders (VAE) [95], or a combination of both [96]). These models are powerful tools for producing new samples, but are not widely used for extending human full-body datasets for three reasons: 1) Performance of the generative models is still unsatisfactory, 2) The generated data is unlabelled, 3) Not only learning high-quality generative models for human full-body takes too much time, but it also requires large amount of high-resolution learning data, which is yet not available. Therefore, researchers [97], [64], [98], [99], [100], [82], [30] mostly either perform transfer learning to capture the useful knowledge of large datasets or resort to the simple yet useful label-persevering augmentation techniques from basic data augmentation (flipping, shifting, scaling, cropping, resizing, rotating, shearing, zooming, etc.) to more sophisticated methods such as random erasing [101] and foreground augmentation [102].

*Data Augmentation*. In this context, Bekele et al. [97] studied the effectiveness of 3 basic data augmentation techniques on their proposed solution and observed that the $F1$ score is improved from 85.7 to 86.4 for an experiment on the PETA dataset. Further, [43] discussed that ResNets could take advantage of the skipped connections to avoid overfitting. Their experimental results on the PETA dataset confirm the superiority of ResNets without augmentation over the SVM-based and plain CNN models.

*Transfer Learning*. In clothing attribute recognition, some works may deal with two domains (types of images): 1) in shop images that are high-quality in specific poses; 2) in-the-wild images that vary in the pose, illumination, and resolution. To address the problem of limited labeled data, we can transfer the knowledge of one domain to the other domain. In this context, inspired by curriculum learning, Dong et al. [44] suggest a two-step framework for curriculum transfer of knowledge from shop clothing images to in-the-wild *similar* clothing images. To this end, they train a multi-task network with easy samples (in-shop) and copy its weights to a triplet-branch curriculum transfer network. At first, these branches have identical weights; however, in the second training stage (with harder examples), the feature similarity values between the target and the positive branches become larger than between the target and negative branches. The ablation studies confirm the effectiveness of the authors' idea and show that the mean average ($mA$) improved from 51.4 to 58.8 for plain multi-task and proposed model, respectively, on the Cross-Domain clothing dataset [103]. Moreover, this work indicates that curriculum learning versus end-to-end learning achieves better results, with 62.3 and 64.4 of $mA$, respectively.

### 2.3. Attributes Relationship

Both the spatial and semantic relationships among the attributes affect the performance of the PAR models. For example, hairstyle and footwear are correlated, while related to different regions (i.e., spatial distributions) of the input data. Regarding the semantic relationship, pedestrian attributes may either conflict with each other or are mutually confirming. For instance, wearing jeans and a skirt is an unexpected outfit, while wearing a T-shirt and sports shoes may co-appear with high probability. Therefore, taking these intuitive interpretations into account could be considered as a refinement step that improves the prediction-list of the attributes [55]. Furthermore, considering the contextual relation between various regions improve the performance of the PAR models. To consider

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



the correlation among the attributes there are several possibilities such as using multi-task architecture [89], multi-label classification with weighted loss function [104], Recurrent Neural Networks (RNN) [51], Graph Convolutional Network (GCN) [56]. We have classified them into two main groups:

- Network-Oriented methods that take advantage of the various implementation of convolutional layers/blocks to discover the relation between attributes,
- math-oriented methods that may or may not extract the features using CNNs, but perform some mathematical operations on the features to modify them regarding the existing intrinsic correlations among the attributes.

In the following, we discuss the literature of both categories.

#### 2.3.1. Network-Oriented Attribute Correlation Consideration

*A) Multi-task Learning.* Network-Oriented Attribute Correlation ConsiderationIn [9], Lu et al. discuss that the intuition-based design of multi-task models is not an optimal solution for sharing the relevant information over multiple tasks, and they propose to gradually widen the structure of the model (add new branches) using an iterative algorithm. Consequently, in the final architecture, correlated tasks share most of the convolutional blocks together, while uncorrelated tasks will use different branches. Evaluation of the model on the fashion dataset [86] shows that by widening the network to 32 branches, the accuracy of the model cannot compete with other counterparts; however, the speed increases (from 34$ms$ to 10$ms$) and the number of parameters decreases from 134 million to 10.5 million.

In a multi-task attribute recognition problem, each task may have a different convergence rate. To alleviate this problem and jointly learn multiple tasks, [45] proposes a weighted loss function that updates the weights for each task in the course of learning. The experimental evaluation on the Market-1501 dataset [17] shows an improvement in accuracy from 86.8% to 88.5%.

In [47] and [48], the authors study the multi-task nature of PAR and attempt to build an optimal grouping of the correlated tasks, based on which they share the knowledge between tasks. The intuition is that, similar to the human brain, the model should learn more manageable tasks first and then uses them for solving more complex tasks. The authors claim that learning correlated tasks needs less effort, while uncorrelated tasks require specific feature representations. Therefore, they apply a curriculum learning schedule to transfer the knowledge of the easier tasks (strongly correlated) to the harder ones (weakly correlated). The baseline results show that learning the tasks individually yields 71.0% accuracy on the SoBiR dataset [105], while this number for learning multiple tasks at once is 71.3% and for a curriculum-based multi-task model is 74.2%.

Considering HAR as a multi-task problem, [8] proposes to improve the model architecture in terms of feature sharing between tasks. Authors claim that by learning a linear combination of features, the inter-dependency of the channels is ignored, and the model cannot exchange spatial information. Therefore, after each convolutional block in the model, they insert a shared module between tasks to share the information. This module considers three aspects: 1) fusing the features of each two tasks together, 2) generating attention maps regarding the location of the attributes [106], and 3) keeping the effect of the original features of each task. Ablation studies over this module's positioning indicate that adding it at the end of the convolutional blocks yields the best results. However, the performance is approximately stable when different branches of the module (one at a time) are ablated.

*B) RNN.* In [50], authors discuss that person re-id focuses on the global features, while attribute recognition relies on local aspects of individuals. Therefore, Liu et al. [50] propose a network consisted of three parts that work together to learn the person's attributes and re-identification (re-id). Further, to capture the contextual spatial relationships and focus to the location of each attribute, they use the RNN-CNN backbone feature extractor followed by an attention model.

To mine the relation of attributes, [52] uses a model based on Long Short Term Memory (LSTM). Intuitively, using several successive stages of LSTM preserves the necessary information along the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



pipeline and forgets the uncorrelated features. In this work, the authors first detect three-body pose-lets based on the skeleton information. They consider the full-body as another pose-let followed by several fully connected layers to produce several groups of features (for each attribute, one group of features). Each group of features is passed to an LSTM block, followed by a fully-connected layer. Finally, the concatenation of all features is considered as the final feature representation of the input image. Considering that LSTM blocks are successively connected to each other, they carry the useful information of previous groups of features to the next LSTM. The ablation study in this work shows that the plain Inception-v3 on PETA dataset attains 85.7 of $F1$ metric, and adding LSTM blocks on top of the baseline improves its performance to 86.0, while the full version of the model that processes the body-parts achieves to $F1$ 86.5.

Regarding the functionality of RNN in contextual combinations in the sequenced data, [53] introduces two different methods to localize the semantic attributes and capture their correlations implicitly. In the first method, the input image's extracted features are divided into several groups; then, each group of features is given to an LSTM layer followed by a regular convolution block and a fully connected layer, while all the LSTM layers are connected together successively. In the second method, all the extracted features from the backbone are multiplied (spatial point-wise multiplication) by the last convolution block's output to provide global attention. The experiments show that dividing the features into groups from global to local features yields better results than random selection.

Inspired by image-captioning methods, [49] introduced a Neural PAR that converts attributes recognition to the image-captioning task. To this end, they generated sentence vectors to describe each pedestrian image using a random combination of attribute-words. However, there are two major disruptions in designing an image-caption architecture for attribute classification: 1) variable length of sentences (attribute-words) for different pedestrians and 2) finding relevance between attributes vectors and spatial space. To address these challenges, the authors used RNNs units and lookup-table, respectively. how much they improved the results in comparison with a plain network? how they implemented this idea?

To deal with low-resolution images, Wang et al. [51] formulated the PAR task as a sequential prediction problem, in which a two-step model is used to encode and decode the attributes for discovering both the context of intra-individual attributes and the inter-attribute relation. To this end, Wang et al. took advantage of LSTMs in both encode and decode steps for different purposes, such that in the encoding step the context of the intra-person attributes is learned, while in the decoding step, LSTMs is utilized to learn the inter-attributes correlation and predict the attributes as a sequence prediction problem. how much they improved the results in comparison with a plain network?

*C) GCN.* In [56], Li et al. introduce a sequential-based model that relies on two graph convolutional networks, in which the semantic attributes are used as the nodes of the first graph, and patches of the input image are used as the nodes of the second graph. To discover the correlation between regions and semantic attributes, they embedded the output of the first graph as the extra inputs into the second graph and vise versa (the output of the second graph is embedded as the extra inputs into the first graph). To avoid a closed loop in the architecture, they defined two separate feed-forward branches, such that the first branch receives the image patches and presents the spatial context representation of them. This representation is then mapped into the semantic space to produce the features that capture the similarity between regions. The second branch input is semantic attributes that are processed using a graph network and mapped into spatial graphs to capture the semantic-aware features. The output of both branches is fused to let and end-to-end learning. The ablation studies show that in comparison with a plain ResNet50 network, the $F1$ results could improve by margins of 3.5 and 1.3 for the PETA and RAP datasets, respectively.

Inspired by [56], in [55], Li et al. present a GCN-based model to yield the human parsing alongside the human attributes. Therefore, a graph is built upon the image features so that each group of features corresponds to one node of the graph. Afterward, to capture the relationships among the groups of attributes, a graph convolution is performed. Finally, for each node, a classifier is learned to predict

 doi:10.20944/preprints202007.0055.v1



the attributes. To produce the human parsing results, they apply a residual block that uses both the original features and the output of the graph convolution in the previous branch. Based on the ablation study, a plain ResNet50 on the PETA dataset achieves a $F1$ score of 85.0, while a model based on body parts yields a $F1$ score of 84.4, and this number for the model equipped with the above-mentioned idea is 87.9.

Tan et al. [57] observed the close relationship between some of the human attributes and claimed that in multi-task architectures, the final loss function layer is the critical point of learning, which may not have sufficient influence for obtaining a comprehensive representation for explaining the attribute correlations. Moreover, the limitation in receptive fields of CNNs [107] hinders the model's ability to effectively learn the contextual relations in the data. Therefore, to capture the structural connections among attributes and contextual information, the authors use two Graph Convolutional Networks (GCN) [108]. However, as image data is not originally structured as graphs, they use the extracted attribute-specific features (each feature corresponds to one attribute) from a ResNet backbone to construct the first graph. For the second graph, clusters of regions (pixels) in the input image are considered as the network nodes. The clusters are learned using the share ResNet backbone –with the previous graph). Finally, the outputs of both graph-based branches are averaged. As LSTM also considers the relationship between parts, authors have replaced their proposed GCNs with LSTMs in the model and observed a slight drop in the model's performance. The ablation strides on three pedestrian datasets show that the $F1$ metric performance of a vanilla model improves with a margin of 2.

[54] recognized the clothing style by mixing extracted features from the body parts. They applied a graph-based model with Conditional Random Fields (CRFs) to explore the correlation between clothes attributes. Specifically, using the weighted sum of body-part features, they trained an SVM for each of the attributes and used CRF to learn the relationships between attributes. By training the CRF with output probability scores from SVM classifiers, the attributes' relationship is explored. Although using CRFs was successful in this work, there are yet some disadvantages: a) due to extensive computational cost, CRFs is not an appropriate solution when a broad set of attributes are considered, and b) CRFs cannot capture the spatial relation between attributes[56] c) models can not simultaneously optimize classifiers and CRFs [56], so it is not useful in an end-to-end model.

#### 2.3.2. Math-Oriented Attribute Correlation Consideration

*A) Grammar.* In [58], Park et al. addressed the need for an interpretable model that can jointly yield the body-pose information (body joints coordinates) and human semantic attributes. To this end, authors implemented an and-or grammar model, in which they integrated three types of grammars: 1) simple grammars that break down the full-body into smaller nodes; 2) dependency grammar that indicates which nodes (body parts) are connected to each other and models the geometric articulations; 3) attribute grammar that assigns the attributes to each node. The ablation studies for attribute prediction showed that the performance is better if the best pose estimation for each attribute is used for predicting the corresponding attribute score.

*B) Multiplication.* In [59], authors discussed that a plain CNN could not handle human multi-attribute classifications effectively, as for each image, several labels have been entangled. To address this challenge, Han et al. [59] proposed to use a ResNet50 backbone followed by multiple branches to predict the occurrence probability of each attribute. Further, to improve the results, they provided a matrix from ground truth labels to obtain the conditional probability of each label (semantic attribute) given another attribute. The multiplication of this matrix by the previously obtained probability provides the models with a priori knowledge about the correlation of attributes. The ablation study indicated that the baseline (plain ResNet50) on the PETA dataset achieves 85.8 of $F1$ metric, while this number for a simple multi-branch model and full-version model is 86.6 and 87.6, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



In order to mitigate the correlation between the visual appearance and the semantic attributes, [60] uses a fusion attention mechanism and provides a balanced-weight between the image-guided and attribute-guided features. First, attributes are embedded in a latent space with the same dimension of the image features. Next, a nonlinear function is applied to the image features to obtain its feature distribution. Then, the image-guided features are obtained via an element-wise multiplication between the feature distribution of the image and the embedded attribute features. To obtain the attribute-guided features, they embed the attributes to a new latent space; next, the results of the element-wise multiplication between image features and embedded attribute features are considered as the input of a nonlinear function, for which its output provides attribute-guided features. Meanwhile, to consider the class imbalance, authors use the focal loss function to train the model. The ablation study shows that the $F1$ metric performance of the baseline on the PETA dataset is 85.6, which improves to 85.9 when the model is equipped with the above-mentioned idea.

In [46], authors propose a multi-task architecture, in which each attribute corresponds to one separate task. However, to consider the relationship between attributes, both the input image and category information are projected into another space, where the latent factors are disentangled. By applying the element-wise multiplication between the feature representation of the image and its class information, the authors define a discriminant function. When using it, a logistic regression model can learn all the attributes simultaneously. To show the efficiency of the methods, authors evaluate their proposed approach in several attribute datasets of animals, objects, and birds.

*C) Loss Function.* Li et al. [61] discussed the attribute relationships and introduced two models to demonstrate the effectiveness of their idea. Considering HAR as a binary classification problem, the authors constructed a plain multi-label CNN that predicts all the attributes at-once. They also equipped the previous model with a weighted-loss function (cross-entropy), in which each attribute classifier has a specific weight to update the network weights for the next epoch. The experimental results on the PETA dataset with 35 attributes indicated that weighted cross-entropy loss function could improve the accuracy prediction in 28 attributes and increase the $mA$ by 1.3 percent.

### *2.4. Occlusion*

In HAR, occlusion is a primary challenge, in which parts of the useful information of the input data may be covered with other subjects/objects. As this situation is likely to occur in real-world scenarios, it is necessary to be handled. In the context of person re-id, [109] claims that constructing the occluded body parts could improve the results, and in the HAR context, [69] suggests that using sequences of pedestrian images somehow alleviates the occlusion problem.

Considering the low-resolution images and partial occlusion of the pedestrian's body, [68] proposed to manipulate the dataset with occurring frequent partial occlusions and degraded the resolution of the data. Then, the authors trained a model to reconstruct the images with high resolution and do not suffer from occlusion. This way, the reconstruction model will help to manipulate the original dataset before training a classification model. As reconstruction is performed with a GAN, the generated images are different from the original annotated dataset and somehow lost part of the annotations, which degrade the overall performance of the system compared to when one uses the original dataset for training. However, the ablation study in this paper shows that if two identical classification networks are separately trained on corrupted and reconstructed data, the performance of the model that learns from the reconstructed data is better with a high margin.

To tackle the problem of occlusion, [67] proposes to use a sequence of frames for recognizing human attributes. First, they extract the frame-level spatial features using a shared ResNet-50 backbone feature extractor [110]. The extracted features are then processed in two separate paths, one of them learns the body pose and motion, and the other branch learns the semantic attributes. Finally, each attribute's classifier uses an attention module that generates an attention vector showing the importance of each frame for attribute recognition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



To address the challenge of partial occlusion, [98] and [69] adopted video datasets for attributes recognition as often occlusions are a temporary situation. [98] divided each video clip to several pieces and extracted a random frame from each piece to create a new video clip with a few frame length. The final recognition confidence of each attribute is obtained by aggregating the recognition probability on the selected frames.

### *2.5. Classes Imbalance*

The existence of large differences between the number of samples for each attribute (class) is known as data class imbalance. Generally, in multi-class classification problems, the ideal scenario would be to use the same amount of data for each class, in order to preserve the learning importance of all the classes at the same level. However, the classes in HAR datasets are naturally imbalanced since the number of samples of some attributes (e.g., wearing skirts) are lower than others (e.g., wearing jeans). Large class imbalance causes over-fitting in classes with limited data, while classes with a large number of samples need more training epochs to converge. To address this challenge, some methods attempt to balance the number of samples in each class as a pre-processing step [111], [112], [113], which are called *hard solutions*. Hard solutions are classified into three groups: 1) up-sampling the minority classes, 2) down-sampling the large classes, and 3) generating new samples. On the other hand, *soft solutions* are interested in handling the data class imbalance by introducing new training methods [65] or novel loss functions, in which the importance of each class is weighted based on the frequencies of the data [114], [115], [116]. Furthermore, the combination of both solutions has been the subject of some studies [66].

#### 2.5.1. *Hard* Solutions

The earlier hard solutions are focused either on interpolation between the samples [117], [118], or clustering the dataset and oversampling by cluster-based methods [119]. The primary way of up-sampling in deep learning is to augment the existing samples –as discussed in section 2.2. However, excessive up-sampling may lead to over-fitting when the classes are highly imbalanced. Therefore, some works down-sample the majority classes [120]. Random down-sampling may be an easy choice, but [121] proposes to use the boundaries among the classes to remove redundant samples. However, loss of information is an inevitable part of down-sampling, as some samples are removed, which may carry useful information.

To address these problems, Fukui et al. [62] designed a multi-task CNN, in which classes (attributes) with fewer samples are given more importance in the learning phase. The batch of samples in conventional learning methods are selected randomly; therefore, the rare examples are less likely to be in the mini-batch. Meanwhile, data augmentation cannot be sufficient for balancing the dataset as ordinary data augmentation techniques generate new samples regardless of their rarity. Therefore, Fukui et al. [62] defines a rarity rate for each sample in the dataset and perform the augmentation for rare samples. Later, from the created mini-batches, those with appropriate sample balance are selected for training the model. The experimental results on a dataset with four attributes show a slight improvement in the average recognition rate, though the superiority is not consistent for all the attributes.

#### 2.5.2. Soft Solutions

As previously mentioned, soft solutions focus on boosting the learning methods' performance, rather than merely increasing/decreasing the number of samples. Designing loss functions is a popular approach for guiding the model to take full advantage of the minority samples. For instance, [60] proposes the combination of focal loss [122] and cross-entropy loss functions to introduce a focal cross-entropy loss function (see section 2.3.2 for the analytical review over [60]).

Considering the success of curriculum learning [123] in other fields of studies, in [65], the author addressed the challenge of imbalance-distributed data in HAR by batch-based adjustment of data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



sampling strategy and loss weights. It was argued that providing balanced distribution from a highly imbalanced dataset (using sampling strategies) for the whole learning process may cause the model to disregard the samples with most variations (i.e., classes with majority samples) and only emphasizes on the minority class. Moreover, the weighted terms in loss functions play an essential role in the learning process. Therefore, both the classification loss (often cross-entropy) and metric learning loss (which aims to learn feature embedding for distinguishing between samples) should be handled based on their importance. To consider these aspects, authors defined two schedules, one for adjusting the sampling strategy by re-ordering the data from imbalanced to balanced and easy to hard; and the other curriculum schedule handles the loss importance between classification and distance metric learning. The ablation study in this work showed that the sampling scheduler could increase the results of a baseline model form 81.17 to 86.58, and adding loss scheduler to it could improve the results to 89.05.

To handle the class imbalance problem, [63] modifies the focal loss function [122] and apply it for an attention-based model to focus on the hard samples. The main idea is to add a scaling factor to the binary cross-entropy loss function to down-weight the effect of easy samples with high confidence. Therefore, the hard misclassified samples of each attribute (class) add larger values to the loss function and become more critical. Considering the usual weakness of attention mechanism that does not consider the location of an attribute, the authors modified the attention masks in multiple levels of the model using attribute confidence weighting. Their ablation studies on the WIDER dataset [26] with ResNet-101 backbone feature extractor [110] showed the plain model achieves mA 83.7 and applying the weighted focal loss function improve the results to 84.4 while adding the multi-scale attention increased it to 85.9.

#### 2.5.3. Hybrid Solutions

Hybrid approaches use the combination of the above-mentioned techniques. Performing data augmentation over the minority classes and applying a weighted loss function or a curriculum learning strategy are examples of hybrid solutions for handling the class data imbalance. In [66], the authors discuss that learning from an unbalanced dataset leads to biased classification, with higher classification accuracy over the majority classes and lower performance over the minority classes. To address this issue, Chawla et al. [66] proposed an algorithm that focuses on difficult samples (misclassified). To implement this strategy, the authors took advantage of [117], which generates new synthetic instances in each training iteration from the minority classes. Consequently, the weights for the minority samples (false negatives) are increased, which improves the model's performance.

### *2.6. Part-Based And Attribute Correlation-Based Methods*

"Whether considering a group of attributes together improve the results of an attribute recognition model or not?" is the question that [124] tries to answer by addressing the correlation between attributes using a Conditional Random Field (CRF) strategy. Concerning the calculated probability distribution over each attribute, all the Maximum A Posterioris (MAPs) are estimated, and then, the model searches for the most probable mixture in the input image. To also consider the location of each attribute, authors extract the part patches based on the bounding box around the full-body, as in fashion datasets pose variations are not significant. A comparison between several simple baselines shows that the CRF-based method (0.516$F1$ score) works slightly better than a localization-based CNN (0.512$F1$ score) on the Chictopia dataset [125], while a global-based CNN $F1$ performance is 0.464.

### *2.7. Category-Based Performance Comparison*

Table 1 shows the performance of the HAR approaches over the last decade and indicates a consistent improvement of methods over time. In 2016, the performance evaluation of [29] on the RAP and PETA datasets achieved to $F1$ score 66.12 and 84.90, respectively, while these number were improved to 79.98 and 86.87 in the year 2019 [38]. Furthermore, according to Table 1, it is clear that challenges of attributes localization and attributes correlation have attracted the most attention over the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



recent years, which indicates that extracting distinctive fine-grained features from relevant locations of the given input images is the most important aspect of HAR models.

Despite the early works that analyzed the human full-body data in different locations and situations, recent works have focused on attribute recognition from surveillance data, which arouses some privacy issues.

Appearing comprehensive evaluation metrics is another noticeable change over the last decade. Due to the intrinsic, large class imbalance in the HAR datasets, $mA$ cannot provide a comprehensive performance evaluation over different methods. Suppose that in a binary classification situation, if 99% of the samples belong to persons with glasses and 1% of samples belong to persons without glasses, the model can recognize all the test samples as persons with glasses and still has 99% of accuracy in recognition. Therefore, for a fair performance comparison with the state of the arts, it is necessary to consider metrics such as $Prec$, $Rec$, $Acc$, and $F1$ – which are discussed in section 4.

Table 1 also shows that RAP and PETA datasets have attracted the most attention in the context of attribute recognition –which excludes person re-id.

Last but not least, we observe that the performance of the state of the arts is yet far from the reliable range to be used in forensic affairs and enterprises, and it requires more attention in both introducing novel datasets and proposing robust methods.

**Table 1.** Performance comparison of HAR approaches over the last decade for different benchmarks.

| Ref. | Taxonomy | Dataset | mA | Acc. | prec. | rec. | F1 |
|---|---|---|---|---|---|---|---|
| [27], 2011 | Pose-Let | HAT [27] | 53.80 | - | - | - | - |
| [24], 2011 | Pose-Let | [24] | 82.90 | - | - | - | - |
| [54], 2012 | GCN | [54] | - | 84.90 | - | - | - |
| [35], 2013 | Body-Part | HAT [27] | 69.88 | - | - | - | - |
| [25], 2013 | Pose-Let | HAT [27] | 59.30 | - | - | - | - |
| [28], 2013 | Pose-Let | HAT [27] | 59.70 | - | - | - | - |
| [33], 2015 | Body-Part | DP [24] | 83.60 | - | - | - | - |
| [33], 2015 | Body-Part | DP [24] | 83.60 | - | - | - | - |
| [26], 2016 | Pose-Let | WIDER [26] | 92.20 | - | - | - | - |
| [29], 2016 | Pose-Let | RAP [104] | 81.25 | 50.30 | 57.17 | 78.39 | 66.12 |
| | | PETA [126] | 85.50 | 76.98 | 84.07 | 85.78 | 84.90 |
| [36], 2017 | Attention | WIDER [26] | 82.90 | - | - | - | - |
| | | Berkeley [24] | 92.20 | - | - | - | - |
| [37], 2017 | Attention | RAP [104] | 76.12 | 65.39 | 77.33 | 78.79 | 78.05 |
| | | PETA [126] | 81.77 | 76.13 | 84.92 | 83.24 | 84.07 |
| | | PA-100K [37] | 74.21 | 72.19 | 82.97 | 82.09 | 82.53 |
| [72], 2017 | Pose Estimation | RAP [104] | 77.70 | 67.35 | 79.51 | 79.67 | 79.59 |
| | | PETA [126] | 83.45 | 77.73 | 86.18 | 84.81 | 85.49 |
| | | WIDER [26] | 82.40 | - | - | - | - |
| [36], 2017 | Attention | RAP [104] | 78.68 | 68.00 | 80.36 | 79.82 | 80.09 |
| | | PA-100K [37] | 76.96 | 75.55 | 86.99 | 83.17 | 85.04 |
| [43], 2017 | Loss Function | PETA [126] | - | 75.43 | - | 70.83 | - |
| [44], 2017 | Limited Data | [44] | 64.35 | - | 64.97 | 75.66 | - |
| [45], 2017 | Multitask | Market [17] | - | 88.49 | - | - | - |
| | | Duke [126] | - | 87.53 | - | - | - |
| [9], 2017 | Multitask | D.Fashion | - | 83.24* | - | - | - |
| [127], 2018 | Pose Estimation | PETA [126] | 82.97 | 78.08 | 86.86 | 84.68 | 85.76 |
| | | RAP [104] | 74.31 | 64.57 | 78.86 | 75.90 | 77.35 |

*Continued on next page*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



Table 1 – *Continued from previous page*

| Ref. | Category | Dataset | mA | Acc. | prec. | rec. | F1 |
|---|---|---|---|---|---|---|---|
| | | PA-100K [37] | 74.95 | 73.08 | 84.36 | 82.24 | 83.29 |
| [42], 2019 | Attribute Location | RAP [104] | 81.87 | 68.17 | 74.71 | 86.48 | 80.16 |
| | | PETA [126] | 86.30 | 79.52 | 85.65 | 88.09 | 86.85 |
| | | PA-100K [37] | 80.68 | 77.08 | 84.21 | 88.84 | 86.46 |
| [38], 2019 | Attention | PA-100K [37] | 81.61 | 78.89 | 86.83 | 87.73 | 87.27 |
| | | RAP [104] | 81.25 | 67.91 | 78.56 | 81.45 | 79.98 |
| | | PETA [126] | 84.88 | 79.46 | 87.42 | 86.33 | 86.87 |
| | | Market [17] | 87.88 | - | - | - | - |
| | | Duke[126] | 87.88 | - | - | - | - |
| [40], 2019 | Attention | RAP [104] | 84.28 | 59.84 | 66.50 | 84.13 | 74.28 |
| | | WIDER [26] | 88.00 | - | - | - | - |

## 3. Datasets

As opposed to other surveys, instead of merely enumerating the datasets, in this manuscript, we discuss the advantages and drawbacks of each dataset, with emphasis on data collection methods/software. Finally, we discuss the intrinsically imbalanced nature of HAR datasets and other challenges that arise when gathering data.

PETA dataset

**PE**des**T**rian **A**ttribute (PETA) [126] dataset combines 19000 pedestrian images gathered from 10 publicly available datasets; therefore the images present large variations in terms of scene, lighting conditions and image resolution. The resolution of the images varies from 17 $\times$ 39 to 169 $\times$ 365 pixels. The dataset provides rich annotations: the images are manually labeled with 61 binary and 4 multi-class attributes. The binary attributes include information about demographics (gender: *Male*, age: *Age16-30, Age31-45,Age46-60, AgeAbove61*), appearance (*long hair*), clothing (*T-shirt, Trousers* etc.) and accessories (*Sunglasses, Hat, Backpack* etc.). The multi-class attributes are related to (eleven basic) color(s) for the upper-body and lower-body clothing, shoe-wear, and hair of the subject. When gathering the dataset, the authors tried to balance the binary attributes; in their convention, a binary class is considered balanced if the maximal and minimal class ratio is less than 20:1. In the final version of the dataset, more than half of the binary attributes (31 attributes) have a balanced distribution.

RAP dataset

Currently, there are two versions of the RAP (**R**ichly **A**nnotated **P**edestrian) dataset. The first version, RAP-v1 v1 [128] was collected from a surveillance camera in shopping malls over a period of three months; next, 17 hours of video footage were manually selected for attribute annotation. In total, the dataset comprises 41585 annotated human silhouettes. The 72 attributes labeled in this dataset include demographic information (*gender* and *age*), accessories (*backpack, single shoulder bag, handbag, plastic bag, paper bag* etc.), human appearance (*hair style, hair color, body shape*) and clothing information (*clothes style, clothes color, footware style, footware color* etc.). In addition, the dataset provides annotations about occlusions, viewpoints and body-parts information.

The second version of the RAP dataset [104] is intended as a unifying benchmark for both person retrieval and person attribute recognition in real-world surveillance scenarios. The dataset was captured indoor, in a shopping mall and contains 84928 images (2589 person identities) from 25 different scenes. High-resolution cameras (1280 $\times$ 720) were used to gather the dataset, and the resolution of human silhouettes varies from 33 $\times$ 81 to 415 $\times$ 583 pixels. The attributes annotated are the same as in RAP v2 (72 attributes, and occlusion, viewpoint, and body-parts information).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



Parse27k dataset

**P**edestrian **a**ttribute **r**ecognition in **se**quences (Parse27k) dataset [129] contains over 27000 pedestrian images, annotated with 10 attributes. The images were captured by a moving camera across a city environment; every $15^{th}$ video frame was fed to the Deformable Part Model(DPM) pedestrian detector [130] and the resulting bounding boxes were annotated with the 10 attributes based on binary or multinomial propositions. As opposed to other datasets, the authors also included an N/A state (i.e. the labeler cannot decide on that attribute). The attributes from this dataset include gender information (3 categories: *male, female, N/A*), accessories (*Bag on Left Shoulder, Bag on Right Shoulder Bag in Left Hand, Bag in Right Hand, Backpack*; each with three possible states: *yes, no, N/A*), orientation (with *4 + N/A* or *8 + N/A* discretizations) and action attributes: *posture* (*standing*, *walking*, *sitting* and *N/A*) and *isPushing* (*yes, no, N/A*). As the images were initially processed by a pedestrian detector, the images of this dataset consist of a fixed-size bounding region of interest, and thus are strongly aligned and contain only a subset of possible human poses.

CRP dataset

CRP (**C**altech **R**oadside **P**edestrians) [131] dataset was captured in real world conditions, from a moving vehicle. The position (bounding-box) of each pedestrian, together with 14 body joints are annotated in each video frame. CRP comprises 4222 video tracks, with 27454 pedestrian bounding boxes. The following attributes are annotated for each pedestrian: age ( 5 categories: *child*, *teen*, *young adult*, *middle aged* and *senior*), gender (2 categories: *female* and *male*), weight (3 categories: *Under*, *Healthy* and *Over*), and clothing style (4 categories: *casual*, *light athletic*, *workout* and *dressy*). The original, un-cropped videos together with the annotations are publicly available.

Describing People dataset

Describing People dataset [24] comprises 8035 images from the H3D [73] and the PASCAL VOC 2010 [132] datasets. The images from this database are aligned, in the sense that for each person, the image is cropped (by leaving some margin) and then scaled so that the distance between the hips and the shoulders is 200 pixels. The dataset features 9 binary (True/False) attributes, as follows: gender (*is male*), appearance (*long hair*), accessories (*glasses*) and several clothing attributes (*has hat, has t-shirt, has shorts, has jeans, long sleeves, long pants*). The dataset was annotated on Amazon Mechanical Turk by five independent labelers; the authors considered a valid label if at least four of the five annotators agreed on its value.

HAT dataset

**H**uman **AT**tributes (HAT) [27,80] contains 9344 images gathered from Flickr; for this purpose, the authors used more than 320 manually specified queries to retrieve images related to people and then, employed an off-the-shelf person detector to crop the humans in the images. The false positives were manually removed. Next, the images were labeled with 27 binary attributes; these attributes incorporate information about the gender (Female), age (Small baby, Small kid, Teen aged, Young (college), Middle Aged, Elderly), clothing (Wearing tank top, Wearing tee shirt, Wearing casual jacket, Formal men suit, Female long skirt, Female short skirt, Wearing short shorts, Low cut top, Female in swim suit, Female wedding dress, Bermuda/beach shorts), pose (Frontal pose, Side pose, Turned Back), Action (Standing Straight, Sitting, Running/Walking, Crouching/bent, Arms bent/crossed) and occlusions (Upper body). The images have high variations both in image size and in the subject's position.

WIDER dataset

WIDER Attribute dataset [26] comprises a subset of 13789 images selected from the WIDER database [133], by discarding the images full of non-human objects and the images in which the human

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



attributes are indistinguishable; the human bounding boxes from these images are annotated with 14 attributes. The images contain multiple humans under different and complex variations. For each image, the authors selected a maximum of 20 bounding boxes (based on their resolution), so in total, there are more than 57524 annotated individuals. The attributes follow a ternary taxonomy: positive, negative and unspecified, and include information about age (*Male*), clothing (*Tshirt, longSleeve, Formal, Shorts, Jeans, Long Pants, Skirt), accessories (Sunglasses, Hat, Face Mask, Logo*), appearance (*Long Hair*). In addition, each image is annotated into one of 30 event classes (meeting, picnic, parade, etc.), thus allowing to correlate the human attributes with the context they were perceived in.

CAD dataset

**C**lothing **A**ttributes **D**ataset [54] uses images gathered from the website Sartorialist[1] and Flikcr. The authors downloaded several images, mostly of pedestrians, and applied an upper-body detector to detect humans; they ended up with 1856 images. Next, the ground truth was established by labelers from Amazon Mechanical Turk. Each image was annotated by 6 independent individuals, and a label was accepted as ground truth if it has at least 5 agreements. The dataset is annotated with the gender of the wearer, information about the accessories (*Wearing scarf, Collar presence, Placket presence*) and with several attributes regarding the clothing appearance (*clothing pattern, major color, clothing category, neckline shape* etc.)

DukeMTMC dataset

**Duke**MTMC-reid (**M**ulti-**T**arget, **M**ulti-**C**amera) dataset [134] was collected in Duke's university campus and contains more than 14 hours of video sequences gathered from 8 cameras, positioned such that they capture crowded scenes. The main purpose of this dataset was person re-identification and multi-camera tracking; however, a subset of this dataset was annotated with human attributes. The annotations were provided at the identity level, and they included 23 attributes, regarding the gender (male, female), accessories: wearing hat (yes, no), carrying a backpack (yes, no), carrying a handbag (yes, no), carrying other types of the bag (yes, no), and clothing style: shoe type (boots, other shoes), the color of shoes (dark, bright), length of upper-body clothing (long, short), 8 colors of upper-body clothing (black, white, red, purple, gray, blue, green, brown) and 7 colors of lower-body clothing (black, white, red, gray, blue, green, brown). Due to violation of civil and human rights, as well as privacy issues, since June 2019, Duke University has terminated the DukeMTMC dataset page.

PA-100K dataset

PA-100k dataset [37] was developed with the intention to surpass the existing HAR datasets both in quantity and in diversity; the dataset contains more than 100000 images captured in 598 different scenarios. The dataset was captured by outdoor surveillance cameras; therefore, the images provide large variance in image resolution, lighting conditions, and environment. The dataset is annotated with 26 attributes, including demographic (age, gender), accessories (handbag, phone) and clothing information.

APiS dataset

The **A**ttributed **P**edestrians **i**n **S**urveillance dataset [135] gathers images from four different sources: KITTI database [136], CBCL Street Scenes [137][2], INRIA database [138] and some video sequences collected by the authors at a train station; in total APiS comprises 3661 images. The human bounding boxes are detected using an off-the-shelf pedestrian detector, and the results are manually processed by the authors: the false positives and the low-resolution images (smaller than

[1] https://www.thesartorialist.com/
[2] http://cbcl.mit.edu/software-datasets/streetscenes/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



90 pixels in height and 35 pixels in width) are discarded. Finally, all the images of the dataset are normalized in the sense that the cropped pedestrian images are scaled to 128×48 pixels. These cropped images are annotated with 11 ternary attributes (positive, negative, and ambiguous) and 2 multi-class attributes. These annotations include demographic (gender) and appearance attributes (*long hair*), as well as information about accessories (*back bag, S-S (Single Shoulder) bag, hand carrying*) and clothing (*shirt, T-shirt, long pants, M-S (Medium and Short) pants, long jeans, skirt, upper-body clothing color, lower-body clothing color*). The multi-class attributes are the two attributes related to the clothing color. The annotation process is performed manually and divided into two stages: annotation stage (the independent labeling of each attribute) and validation stage (which exploits the relationship between the attributes to check the annotation; also, in this stage, the controversial attributes are marked as ambiguous).

Market-1501. dataset

Market-1501 attribute [17,139] dataset is a version of the Market-1501 dataset augmented with the annotation of 27 attributes. Market-1501 was initially intended for cross camera person re-identification, and it was collected outdoor in front of a supermarket using 6 cameras (5 high-resolution cameras and one low resolution). The attributes are provided at the identity level, and in total, there are 1501 annotated identities. In total, the dataset has 32668 bounding boxes for these 1501 identities. The attributes annotated in *Market-1501 attribute* include demographic information (gender and age), information about accessories (*wearing hat, carrying backpack, carrying bag, carrying handbag*), appearance (*hair length*) and clothing type and color (*sleeve length, length of lower-body clothing, type of lower-body clothing, 8 color of upper-body clothing, 9 color of lower-body clothing*).

### *3.1. Fashion Datasets*

DeepFashion dataset

[86] The dataset was gathered from shopping websites, as well as image search engines (blogs, forums, user-generated content). In the first stage, the authors downloaded 1320078 images from shopping websites and 1273150 images from Google images. After a data cleaning process, in which duplicate, out-of-scope, and low-quality images were removed, 800000 clothing images remained to construct the DeepFashion dataset. The images are annotated solely with clothing information; these annotations are divided into categories (50 labels: dress, blouse, etc.) and attributes (1000 labels: adjectives describing the categories). The categories were annotated by expert labelers, while for the attributes, due to their huge number, the authors resorted to meta-data annotation (provided by Google search engine or by the shopping website). In addition, a set of clothing landmarks, as well as their visibility, are provided for each image.

DeepFashion is split into several benchmarks for different purposes: category and attribute prediction (classification of the categories and the attributes), in-shop clothes retrieval (determine if two images belong to the same clothing item), consumer-to-shop clothes retrieval (matching consumer images to their shop counterparts) and fashion landmark detection.

### *3.2. Aerial Datasets*

Over recent years, as their cost has diminished considerably, UAVs applications extended rapidly in various surveillance scenarios. As a response, several UAVs datasets have been collected and made publicly available to the scientific community. Most of them are intended for human detection [143,144], action recognition [145] or re-identification [146].

To the best of our knowledge, the problem P-DESTRE [140] is the first dataset to address the problem of HAR from aerial images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



**Table 2.** Pedestrian attributes datasets.

| Dataset Type | Dataset | #images | Demographic | Accessories | Appearance | Clothing | Colour | Setup |
|---|---|---|---|---|---|---|---|---|
| Pedestrian | PETA[126] | 19000 | ✓ | ✓ | ✓ | ✓ | ✓ | 10 databases |
| | RAP v1 [128] | 41585 | ✓ | ✓ | ✓ | ✓ | ✓ | indoor static camera |
| | RAP v2 [104] | 84928 | ✓ | ✓ | ✓ | ✓ | ✓ | indoor static camera |
| | DukeMTMC† | 34183 | ✓ | ✓ | ✗ | ✓ | ✓ | outdoor static camera |
| | PA-100K[37] | 100000 | ✓ | ✓ | ✗ | ✓ | ✗ | outdoor, surveillance cameras |
| | Market-1501 [17] | 1501 | ✓ | ✓ | ✓ | ✓ | ✓ | outdoor |
| | P-DESTRE [140] | 14M | ✓ | ✓ | ✓ | ✓ | ✗ | UAV |
| Full body | Parse27k [129] | 27000 | ✓ | ✓ | ✗ | ✗ | ✗ | outdoor moving camera |
| | CRP [131] | 27454 | ✓ | ✓ | ✗ | ✗ | ✗ | moving vehicle |
| | APiS [135] | 3661 | ✓ | ✓ | ✓ | ✓ | ✓ | 3 databases |
| | HAT [27] | 9344 | ✓ | ✓ | ✗ | ✓ | ✗ | Flickr |
| | CAD [54] | 1856 | ✓ | ✓ | ✗ | ✓ | ✓ | website crawling |
| | Describing People [24] | 8035 | ✓ | ✓ | ✗ | ✓ | ✗ | 2 databases |
| | WIDER [26] | 13789 | ✓ | ✓ | ✓ | ✓ | ✗ | website crawling |
| Synthetic | CTD [141] | 880 | ✗ | ✗ | ✗ | ✓ | ✓ | generated data |
| | CLOTH3D [142] | 2.1M | ✗ | ✗ | ✗ | ✓ | ✓ | generated data |

† Permanently suspended regarding privacy issues.

P-DESTRE dataset

P-DESTRE dataset [140] was collected in the campuses of two Universities from India and Portugal, using DJI-Phantom 4 drones controlled by human operators. The dataset provides annotations both for person re-identification, as well as for attribute recognition. The identities are consistent across multiple days. The annotations for the attributes include demographic information: *gender, ethnicity* and *age*, appearance information: *height, body volume, hair color, hairstyle, beard, moustache*; accessories information: *glasses, head accessories, body accessories*; *clothing* information and *action* information. In total, the dataset contains over 14 million person bounding boxes, belonging to 261 known identities.

### *3.3. Synthetic Datasets*

Virtual reality systems and synthetic image generation have become prevalent in the last few years, and their results are more and more realistic and of high resolution. Therefore, we also discuss some data sources comprising computer-generated images. It is a well-known fact that the performance of deep learning methods is highly dependent on the amount and distribution of data they were trained on, and synthetic datasets could theoretically be used as an inexhaustible source of diverse and balanced data. In theory, any combination of attributes in any amount could be synthetically generated.

DeepFashion - Fashion image synthesis

The authors of DeepFashion [86] introduce FashionGAN, an adversarial network for generating clothing images on a wearer [147]. FashionGAN is organized into two stages: on a first level, the network generates a semantic segmentation map modeling the wearer's pose. In the second level, a generative model renders an image with precise regions and textures conditioned on this map. In this context, the DeepFashion dataset was extended with 78979 images (taken for the In-shop Clothes Benchmark), associated with several caption sentences and a segmentation map.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



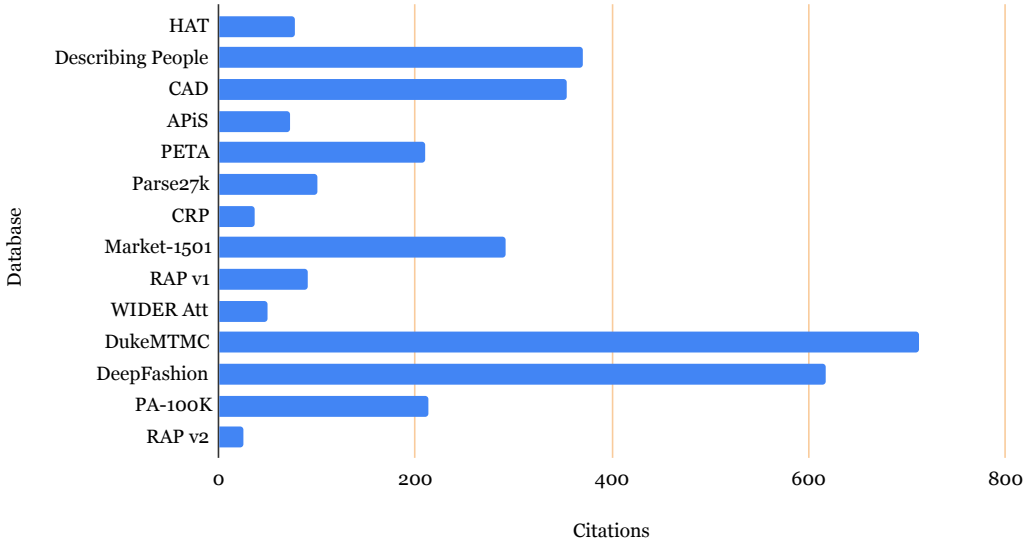


**Figure 3.** HAR datasets citations. The datasets are arranged in increasing order by their publication date. The "oldest" dataset being HAT, published in 2009, while the latest is RAP v2, published in 2018.

CTD dataset

Clothing Tightness dataset [141] (CTD) comprises 880 3D human models, under various poses, both static and dynamic, "dressed" with 228 different outfits. The garments in the dataset are grouped under various categories, such as "T/long shirt, short/long/down coat, hooded jacket, pants, and skirt/dress, ranging from ultra-tight to puffy". CTD was gathered in the context of a deep learning method that maps a 3D human scan into a hybrid geometry image. This synthetic dataset has important implications in virtual try-on systems, soft biometrics, and body pose evaluation. The main drawbacks of this dataset are that it cannot capture exaggerated human postures of low 3D human scans.

CLOTH3D dataset

CLOTH3D [142] comprises thousands of 3D sequences of animated human silhouettes, "dressed" with different garments. The dataset features a large variation on the garment shape, fabric, size, and tightness, as well as human pose. The main applications of this dataset listed by the authors include: "human pose and action recognition in-depth images, garment motion analysis, filling missing vertices of scanned bodies with additional metadata (e.g., garment segments), support designers and animators tasks, or estimating 3D garment from RGB images".

### *3.4. Discussion*

In recent years, HAR has received much interest from the scientific community, with a relatively large number of datasets developed for this purpose; this is also demonstrated by the number of citations. Figure 3 stands as a proof of it. In the past decade, more than 15 databases related to this research field have been published, and most of them received hundreds of citations.

In Table 2 we chose to taxonomize the attributes semantically into demographic attributes (gender, age, ethnicity), appearance attributes (related to the appearance of the subject, such as hairstyle, hair color, weight, etc.), accessory information (which indicate the presence of a certain accessory, such as a hat, handbag, backpack etc.) and clothing attributes (which describe the garments worn by the subjects). In total, we have described 17 datasets, the majority containing over ten thousand images. These datasets can be seen as a continuous effort made by researchers to provide large amounts of varied data required by the latest deep learning neural networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0055.v1



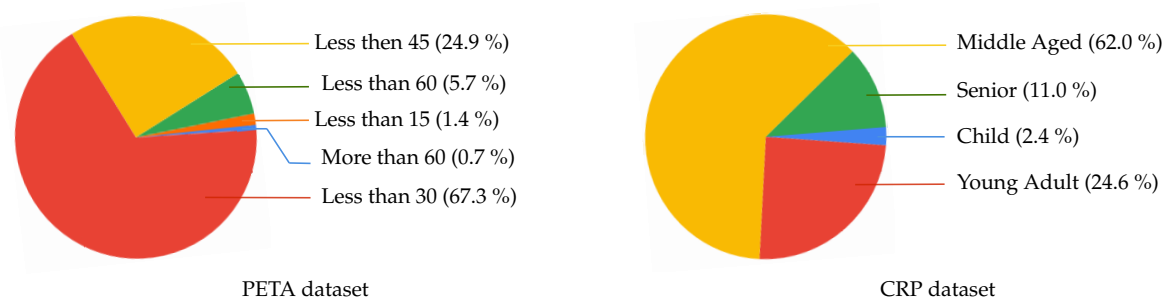


**Figure 4.** Example of the 'Age' distributions for two well-known datasets in the HAR problem.

1. **Attributes definition.** The first issues that should be addressed when developing a new dataset for HAR are: (1) *which attributes should be annotated?* and (2) *how many and which classes are required to describe an attribute properly?*. Obviously, both these questions depend on the application domain of the HAR system. Generally, the ultimate goal on a HAR, regardless of the application domain, would be to accurately describe an image in terms of human-understandable semantic labels, e.g., "a five-year-old boy, dressed in blue jeans, with a yellow T-shirt carrying a striped backpack." As for the second question, the answer is straightforward for some attributes, such as gender, but it becomes more complex and subjective for other attributes, such as age or clothing information. Let's take for example, the age label; different datasets provided different classes for this information: PETA distinguishes between *AgeLess15*, *Age16-30*, *Age31-45*, *Age46-60*, *AgeAbove61*, while CRP dataset adopted a different age classification scheme: *child*, *teen*, *young adult*, *middle aged* and *senior*. Now, if a HAR analyzer is integrated into a surveillance system in a crowded environment, such as Disneyland, and this system should be used to locate a missing child, the age labels from the PETA dataset are not detailed enough, as the "lowest" age class is AgeLess15. Secondly, these differences between the different taxonomies make it difficult to assess the performance of a newly developed algorithm across different datasets.
2. **Unbalanced data.** An important issue in any dataset is related to unbalanced data. Although some datasets were developed by explicitly striking for balanced classes, some classes are not that frequent (especially those related to clothing information), and fully balanced datasets are not a trivial problem. The problem of imbalance also affects the demographic attributes. In all HAR datasets, the class of young children is poorly represented. As an example, in Figure 4 we show the distribution of the age attributes on two popular datasets, CRP and PETA.

   Furthermore, as datasets are usually gathered in a single region (city, country, continent), the data tends to be unbalanced in terms of ethnicity. This is an important issue as some studies [148] proved the existence of *the other race effect* —the tendency to more easily recognize faces from the same ethnicity-– for machine learning classifier.
3. **Data context.** Strongly linked to the problem of data unbalance is the context or environment in which the frames were captured. The environment has a great influence on the distribution of the clothing and demographic (age, gender) attributes. In [26] the authors noticed "strong correlations between image event and the frequent human attributes in it." This is quite logical, as one would expect to encounter more casual outfits in a picnic or sporting event, while at ceremonies (wedding, graduation proms), people tend to be more elegant and dressed-up. The same is valid for the demographic attributes: if the frames are captured in the backyard of a kindergarten, one would that most of the subjects to be children. Ideally, a HAR dataset should provide images captured from multiple and variate scenes. Some datasets explicitly annotated the context in which the data was captured [26], while others address this issue by merging images from various datasets [126]. From another point of view, this leads our discussion to how the images from the datasets are presented. Generally speaking, the dataset provides the images either aligned (all the images have the same size and cropped around the human silhouette with

 doi:10.20944/preprints202007.0055.v1



a predefined margin; e.g., [24]), or make the full video frame/image available and specify the bounding box of each human in the image. We consider that the latter approach is preferable, as it also incorporates context information and allows researches to decide how to handle the input data.

4. **Binary attributes.** Another question in database annotation is what happens when the attribute to annotate is indistinguishable due to low resolution and degraded images, occlusions, or other ambiguities. The majority of datasets tend to ignore this problem and classify the presence of an attribute or provide a multi-class attribute scheme. However, in a real-world setup, we cannot afford this luxury, as the case of indistinguishable attributes might occur quite frequently. Therefore, some datasets [129,135] formulate the attribute classification task with N + 1 classes (+1 for the N/A label). This approach is preferable, as it allows taking both views over the data: depending on the application context, one could simply ignore the N/A attributes or, make the classification problem more interesting, integrate the N/A value into the classification framework.
5. **Camera configuration.** Another aspect that should be taken into account when discussing HAR datasets is the camera setup used to capture the images or video sequences. We can distinguish between fixed-camera and moving-camera setups; obviously, this choice again depends on the application domain into which the HAR system will be integrated. For automotive applications or robotics, one should opt for a moving camera, as the camera movement might influence the visual properties of the human silhouettes. An example of a moving-camera dataset is Parse27k dataset [129]. For surveillance applications, a static camera setup will suffice. In another way, we could distinguish between indoor or outdoor camera setups; for example, RAP dataset [128] uses an indoor camera, while Parse27k dataset [129] comprises outdoor video sequences. Indoor captured datasets, such as [128], although captured in real-world scenarios, do not pose that many challenges as outdoor captured datasets, where the weather and lighting conditions are more volatile. Finally, the last aspect regarding the camera setup is related to the presence of a photographer. If the images are captured by a (professional) photographer some bias is introduced, as a human decides how and when to capture the images, such that it will enhance the appearance of the subject. Some databases, such as CAD [54] or HAT [27,80] use images downloaded from public websites. However, in these images, the persons are aware of being photographed and perhaps even prepared for this (posing for the image, dressed up nicely for a photo session, etc.). Therefore, even if some datasets contain *in-the-wild* images gathered for a different system, they might still contain important differences from *real-world* images in which the subject is unaware of being photographed, the image is captured automatically, without any human intervention, are the subjects are dressed normally and performing natural dynamic movements.
6. **Pose and occlusion labeling.** Another nice to have feature for a HAR dataset is the annotation of pose and occlusions. Some databases already provide this information [27,80,104,128]. Amongst other things, these extra labels prove useful in the evaluation of HAR systems, as they allow researchers to diagnose the errors of HAR and examine the influence of various factors.
7. **Data partitioning strategies.** When dealing with HAR, the datasets partitioning scheme (into the train, validation, and test splits) should be carefully engineered. A common pitfall is to split the frames into the train and validation splits randomly, regardless of the person's identity. This can lead to an unfair assignment of a subject into one of these splits, and inducing bias in the evaluation process. This is even more important, as the current state-of-the-art methods generally rely on deep neural network architectures, which have a black-box behavior in nature, and it is not so straightforward to determine which image features lead to the final classification result.

   Solutions to this problem include extracting each individual (along with its track-lets) from the video sequence or providing the annotations at the identity level. Then, each person could be randomly assigned to one of the dataset splits.

 doi:10.20944/preprints202007.0055.v1



8. **Synthetic data.** Recently, significant advances have been made in the field of computer graphics and synthetic data generation. For example, in the field of drone surveillance, generated data [149] has proven its efficiency in training accurate machine vision systems. In this section, we have presented some computer-generated datasets which contain human attribute annotations. We consider that synthetically generated data is worth taking into consideration, as theoretically, it can be considered an inexhaustible source of data, which could be able to generate subjects with various attributes, under different poses, in diverse scenarios. However, state-of-the-art generative models rely on deep learning, which is known to be "hungry" for data, so data is needed to build a realistic generative model. Therefore, this solution might prove to be just a vicious circle.
9. **Privacy issues.** Last but not least, when gathering a dataset with real-world images, we deal with issues of privacy and human rights violations. Ideally, HAR datasets should contain images captured by real-world surveillance cameras, with the subjects are unaware of being filmed, such that their behavior is as natural as possible. From an ethical perspective, humans should consent before their images are annotated and publicly distributed. However, this is not feasible for all scenarios. As an example, despite its success (if we evaluate success by the number of citations and database downloads), Duke University decided to shut-down Duke-MTMC dataset due to human rights and privacy violation issues.

## 4. Evaluation Metrics

This section reviews the most common metrics used in the evaluation of HAR methods. Considering that HAR is a multi-class classification problem, Accuracy ($Acc$), Precision ($Prec$), Recall ($Rec$), and $F1$ score are the most common metrics for measuring the performance of these methods. In general, these metrics can be calculated at two different levels: label-level and sample-level.

The evaluation at label-level considers each attribute independently. As an example, if the gender and height attributes are considered with the labels (male, female) and (short, medium, high), respectively, the label-level evaluation will measure the performance of each attribute-label combination. The metric adopted in most papers for label-level evaluation is the mean accuracy ($mA$):

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

where $i$ refers to each of the $N$ attributes. $mA$ determines the average accuracy between the positive and negative examples of each attribute.

In the sample-level evaluation, the performance is measured for each attribute disregarding the number of labels that it comprises. $Prec$, $Rec$, $Acc$, and $F1$ score for the $i^{th}$ attribute are thus given by:

$$Prec_i = \frac{TP_i}{P_i}, \quad Rec_i = \frac{TP_i}{N_i}, \quad Acc_i = \frac{TP_i + TN_i}{P_i + N_i}, \quad F_i = \frac{2 * Prec * Rec}{Prec + Rec}. \tag{2}$$

The use of these metrics is very common for providing a comparative analysis of the different attributes. The overall system performance can be either measured by the mean $Acc_i$ over all the attributes or using $mA$. However, these metrics can diverge significantly, when attributes are highly unbalanced. $mA$ is preferred when authors deliberately want to evaluate the effect of data unbalancing.

## 5. Conclusions

This survey reviewed the most relevant works published in the context of human attributes recognition problem (HAR) over the last decade. Contrary to the previous reviews, which provided a methodological categorization of the literature, in this survey we privileged a challenge-based taxonomy, i.e., methods were organized based on the challenges of HAR that they were devised to address. By adopting this type of organization, readers can easily understand the most suitable

strategies for addressing each of the typical challenges of HAR and simultaneously learn which strategies perform better, since for most approaches, we reviewed the accuracy improvement obtained by the proposed strategies. In addition, we comprehensively reviewed the HAR datasets, outlining the advantages and drawbacks of each one, as well as the data collection strategy used. Also, the intrinsically imbalanced nature of the HAR datasets is discussed, as well as the most relevant challenges that typically arise when gathering data for this problem.

**Author Contributions:** Ehsan Yaghoubi's contributions are in conceptualization, methodology, formal analysis, data curation, and writing the original draft. Farhad Khezli collaborated in conceptualization, methodology, and formal analysis. Diana Borza collaborated in writing, conceptualization, data curation, and formal analysis of sections 3 and 1.1 and reviewed the manuscript. SV Aruna Kumar collaborated in formal analysis and in the data curation of section 2.7. João Neves collaborated in reviewing, editing the manuscript, and funding acquisition. Hugo Proença is the supervisor of the project and performed the project administration, conceptualization, funding acquisition, and manuscript revision.

**Funding:** This research is funded by the "FEDER, Fundo de Coesao e Fundo Social Europeu" under the "PT2020 - Portugal 2020" program, "IT: Instituto de Telecomunicações" and "TOMI: City's Best Friend." Also, the work is funded by FCT/MEC through national funds and, when applicable, co-funded by the FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| Acc | Accuracy |
| CAMs | Class Activation Maps |
| CCTV | Closed-Circuit TeleVision |
| CNN | Convolutional Neural Network |
| CRF | Conditional Random Field |
| DPM | Deformable Part Model |
| FP | False Positives |
| GAN | Generative Adversarial Network |
| GCN | Graph Convolutional Network |
| HAR | Human Attribute Recognition |
| re-id | re-identification |
| LMLE | Large Margin Local Embedding |
| LSTM | Long Short Term Memory |
| mA | mean Accuracy |
| MAP | Maximum A Posterioris |
| MAResNet | Multi-Attribute Residual Network |
| Prec | Precision |
| Rec | Recall |
| ResNet | Residual Networks |
| RoI | Regions of Interests |
| RNN | Recurrent Neural Networks |
| SE-Net | Squeeze-and-Excitation Networks |
| SMOTE | Synthetic Minority Over-sampling TEchnique |
| SPR | Spatial Pyramid Representation |
| SSD | Single Shot Detector |
| STN | Spatial Transformer Network |
| SVM | Support Vector Machine |
| TN | True Negative |
| TP | True Positive |
| UAV | Unmanned Aerial Vehicle |
| VAE | Variational Auto-Encoders |
| YOLO | You Only Look Once |

1. Zheng, X.; Guo, Y.; Huang, H.; Li, Y.; He, R. A Survey of Deep Facial Attribute Analysis. *International Journal of Computer Vision* **2020**, pp. 1–33.
2. Wang, X.; Zheng, S.; Yang, R.; Luo, B.; Tang, J. Pedestrian attribute recognition: A survey. *arXiv preprint arXiv:1901.07474* **2019**.
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, pp. 1097–1105.
4. Bekele, E.; Lawson, W. The deeper, the better: Analysis of person attributes recognition. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–8.
5. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* **2016**, *35*, 1285–1298.
6. Alirezazadeh, P.; Yaghoubi, E.; Assunção, E.; Neves, J.C.; Proença, H. Pose Switch-based Convolutional Neural Network for Clothing Analysis in Visual Surveillance Environment. 2019 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2019, pp. 1–5.
7. Yaghoubi, E.; Alirezazadeh, P.; Assunção, E.; Neves, J.C.; Proençaã, H. Region-Based CNNs for Pedestrian Gender Recognition in Visual Surveillance Environments. 2019 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2019, pp. 1–5.
8. Zeng, H.; Ai, H.; Zhuang, Z.; Chen, L. Multi-Task Learning via Co-Attentive Sharing for Pedestrian Attribute Recognition. *arXiv preprint arXiv:2004.03164* **2020**.
9. Lu, Y.; Kumar, A.; Zhai, S.; Cheng, Y.; Javidi, T.; Feris, R. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5334–5343.
10. Masi, I.; Wu, Y.; Hassner, T.; Natarajan, P. Deep face recognition: A survey. 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, 2018, pp. 471–478.
11. Huang, G.B.; Lee, H.; Learned-Miller, E. Learning hierarchical representations for face verification with convolutional deep belief networks. 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2518–2525.
12. Sun, Y.; Liang, D.; Wang, X.; Tang, X. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873* **2015**.
13. De Marsico, M.; Petrosino, A.; Ricciardi, S. Iris recognition through machine learning techniques: A survey. *Pattern Recognition Letters* **2016**, *82*, 106–115.
14. Battistone, F.; Petrosino, A. TGLSTM: A time based graph deep learning approach to gait recognition. *Pattern Recognition Letters* **2019**, *126*, 132–138.
15. Terrier, P. Gait recognition via deep learning of the center-of-pressure trajectory. *Applied Sciences* **2020**, *10*, 774.
16. Layne, R.; Hospedales, T.M.; Gong, S.; Mary, Q. Person re-identification by attributes. Bmvc, 2012, Vol. 2, p. 8.
17. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognition* **2019**, *95*, 151–161.
18. Liu, J.; Kuipers, B.; Savarese, S. Recognizing human actions by attributes. CVPR 2011. IEEE, 2011, pp. 3337–3344.
19. Shao, J.; Kang, K.; Change Loy, C.; Wang, X. Deeply learned attributes for crowded scene understanding. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4657–4666.
20. Tsiamis, N.; Efthymiou, L.; Tsagarakis, K.P. A Comparative Analysis of the Legislation Evolution for Drone Use in OECD Countries. *Drones* **2019**, *3*, 75.
21. Fukui, H.; Yamashita, T.; Yamauchi, Y.; Fujiyoshi, H.; Murase, H. Robust pedestrian attribute recognition for an unbalanced dataset using mini-batch training with rarity rate. 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2016, pp. 322–327.
22. Prabhakar, S.; Pankanti, S.; Jain, A.K. Biometric recognition: Security and privacy concerns. *IEEE security & privacy* **2003**, *1*, 33–42.

23. Dangwei Li, Xiaotang Chen, Z.Z.; Huang, K. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.
24. Bourdev, L.; Maji, S.; Malik, J. Describing people: A poselet-based approach to attribute classification. 2011 International Conference on Computer Vision. IEEE, 2011, pp. 1543–1550.
25. Joo, J.; Wang, S.; Zhu, S.C. Human attribute recognition by rich appearance dictionary. Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 721–728.
26. Li, Y.; Huang, C.; Loy, C.C.; Tang, X. Human attribute recognition by deep hierarchical contexts. European Conference on Computer Vision. Springer, 2016, pp. 684–700.
27. Sharma, G.; Jurie, F. Learning discriminative spatial representation for image classification. BMVC 2011 - British Machine Vision Conference; Hoey, J.; McKenna, S.J.; Trucco, E., Eds.; BMVA Press: Dundee, United Kingdom, 2011; pp. 1–11. doi:10.5244/C.25.6.
28. Sharma, G.; Jurie, F.; Schmid, C. Expanded Parts Model for Human Attribute and Action Recognition in Still Images. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
29. Yu, K.; Leng, B.; Zhang, Z.; Li, D.; Huang, K. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603* **2016**.
30. Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; Bourdev, L. Panda: Pose aligned networks for deep attribute modeling. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1637–1644.
31. Zhu, J.; Liao, S.; Yi, D.; Lei, Z.; Li, S.Z. Multi-label cnn based pedestrian attribute learning for soft biometrics. 2015 International Conference on Biometrics (ICB). IEEE, 2015, pp. 535–540.
32. Zhu, J.; Liao, S.; Lei, Z.; Li, S.Z. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing* **2017**, *58*, 224–229.
33. Gkioxari, G.; Girshick, R.; Malik, J. Actions and attributes from wholes and parts. Proceedings of the IEEE international conference on computer vision, 2015, pp. 2470–2478.
34. Luwei Yang, Ligeng Zhu, Y.W.S.L.; Tan, P. Attribute Recognition from Adaptive Parts. Proceedings of the British Machine Vision Conference (BMVC); Richard C. Wilson, E.R.H.; Smith, W.A.P., Eds. BMVA Press, 2016, pp. 81.1–81.11. doi:10.5244/C.30.81.
35. Zhang, N.; Farrell, R.; Iandola, F.; Darrell, T. Deformable Part Descriptors for Fine-Grained Recognition and Attribute Prediction. The IEEE International Conference on Computer Vision (ICCV), 2013.
36. Guo, H.; Fan, X.; Wang, S. Human attribute recognition by refining attention heat map. *Pattern Recognition Letters* **2017**, *94*, 38–45.
37. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. Proceedings of the IEEE international conference on computer vision, 2017, pp. 350–359.
38. Tan, Z.; Yang, Y.; Wan, J.; Hang, H.; Guo, G.; Li, S.Z. Attention-Based Pedestrian Attribute Analysis. *IEEE transactions on image processing* **2019**, *28*, 6126–6140.
39. Wang, W.; Xu, Y.; Shen, J.; Zhu, S.C. Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
40. Wu, M.; Huang, D.; Guo, Y.; Wang, Y. Distraction-Aware Feature Learning for Human Attribute Recognition via Coarse-to-Fine Attention Mechanism. *arXiv preprint arXiv:1911.11351* **2019**.
41. Liu, P.; Liu, X.; Yan, J.; Shao, J. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102* **2018**.
42. Tang, C.; Sheng, L.; Zhang, Z.; Hu, X. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4997–5006.
43. Bekele, E.; Narber, C.; Lawson, W. Multi-attribute residual network (MAResNet) for soft-biometrics recognition in surveillance scenarios. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 386–393.
44. Dong, Q.; Gong, S.; Zhu, X. Multi-task Curriculum Transfer Deep Learning of Clothing Attributes. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 520–529.

45. He, K.; Wang, Z.; Fu, Y.; Feng, R.; Jiang, Y.G.; Xue, X. Adaptively weighted multi-task deep network for person attribute classification. Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1636–1644.
46. Liang, K.; Chang, H.; Shan, S.; Chen, X. A Unified Multiplicative Framework for Attribute Learning. The IEEE International Conference on Computer Vision (ICCV), 2015.
47. Sarafianos, N.; Giannakopoulos, T.; Nikou, C.; Kakadiaris, I.A. Curriculum learning for multi-task classification of visual attributes. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2608–2615.
48. Sarafianos, N.; Giannakopoulos, T.; Nikou, C.; Kakadiaris, I.A. Curriculum learning of visual attribute clusters for multi-task classification. *Pattern Recognition* **2018**, *80*, 94–108.
49. Ji, Z.; Zheng, W.; Pang, Y. Deep pedestrian attribute recognition based on LSTM. 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 151–155.
50. Liu, H.; Wu, J.; Jiang, J.; Qi, M.; Ren, B. Sequence-based person attribute recognition with joint CTC-attention model. *arXiv preprint arXiv:1811.08115* **2018**.
51. Wang, J.; Zhu, X.; Gong, S.; Li, W. Attribute recognition by joint recurrent learning of context and correlation. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 531–540.
52. Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; Jin, X. Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning. IJCAI, 2018, pp. 3177–3183.
53. Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; Yan, C. Recurrent attention model for pedestrian attribute recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 9275–9282.
54. Chen, H.; Gallagher, A.; Girod, B. Describing clothing by semantic attributes. European conference on computer vision. Springer, 2012, pp. 609–623.
55. Li, Q.; Zhao, X.; He, R.; Huang, K. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 2019, pp. 833–839.
56. Li, Q.; Zhao, X.; He, R.; Huang, K. Visual-semantic graph reasoning for pedestrian attribute recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 8634–8641.
57. Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; Li, S.Z. Relation-Aware Pedestrian Attribute Recognition with Graph Convolutional Networks. AAAI, 2020, pp. 12055–12062.
58. Park, S.; Nie, B.X.; Zhu, S. Attribute And-Or Grammar for Joint Parsing of Human Pose, Parts and Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2018**, *40*, 1555–1569.
59. Han, K.; Wang, Y.; Shu, H.; Liu, C.; Xu, C.; Xu, C. Attribute aware pooling for pedestrian attribute recognition. *arXiv preprint arXiv:1907.11837* **2019**.
60. Ji, Z.; He, E.; Wang, H.; Yang, A. Image-attribute reciprocally guided attention network for pedestrian attribute recognition. *Pattern Recognition Letters* **2019**, *120*, 89–95.
61. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015, pp. 111–115.
62. Fukui, H.; Yamashita, T.; Yamauchi, Y.; Fujiyoshi, H.; Murase, H. Robust pedestrian attribute recognition for an unbalanced dataset using mini-batch training with rarity rate. 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2016, pp. 322–327.
63. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 680–697.
64. Tsaipei Wang, Kai-Chen Shu, C.H.C.Y.F.C. On the Effect of Data Imbalance for Multi-Label Pedestrian Attribute Recognition. 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2018, pp. 74–77.
65. Wang, Y.; Gan, W.; Yang, J.; Wu, W.; Yan, J. Dynamic curriculum learning for imbalanced data classification. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5017–5026.
66. Nitesh V Chawla, Aleksandar Lazarevic, L.O.H.; Bowyer, K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *European Conference on Principles of Data Mining and Knowledge Discovery(PKDD)* **2003**.
67. Chen, Z.; Li, A.; Wang, Y. A temporal attentive approach for video-based pedestrian attribute recognition. Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 2019, pp. 209–220.

68. Fabbri, M.; Calderara, S.; Cucchiara, R. Generative adversarial models for people attribute recognition in surveillance. 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2017, pp. 1–6.
69. Xu, J.; Yang, H. Identification of pedestrian attributes based on video sequence. 2018 IEEE International Conference on Advanced Manufacturing (ICAM). IEEE, 2018, pp. 467–470.
70. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.s. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 4913–4922.
71. Krause, J.; Gebru, T.; Deng, J.; Li, L.J.; Fei-Fei, L. Learning features and parts for fine-grained recognition. 2014 22nd International Conference on Pattern Recognition. IEEE, 2014, pp. 26–33.
72. Sarfraz, M.S.; Schumann, A.; Wang, Y.; Stiefelhagen, R. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089* **2017**.
73. Bourdev, L.; Malik, J. Poselets: Body part detectors trained using 3d human pose annotations. 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009, pp. 1365–1372.
74. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
75. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
76. Jaderberg, M.; Simonyan, K.; Zisserman, A.; others. Spatial transformer networks. Advances in neural information processing systems, 2015, pp. 2017–2025.
77. Jaderberg, M.; Simonyan, K.; Zisserman, A.; others. Spatial transformer networks. Advances in neural information processing systems, 2015, pp. 2017–2025.
78. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, Vol. 2, pp. 2169–2178.
79. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
80. Ramanan, P.F.F..R.B.G..D.M..D. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* **2009**, *32*, 1627–1645.
81. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
82. Zhang, Y.; Gu, X.; Tang, J.; Cheng, K.; Tan, S. Part-based attribute-aware network for person re-identification. *IEEE Access* **2019**, *7*, 53585–53595.
83. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. Advances in neural information processing systems, 2014, pp. 487–495.
84. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
85. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. European conference on computer vision. Springer, 2016, pp. 483–499.
86. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1096–1104.
87. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
88. Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; Wang, X. Learning spatial regularization with image-level supervisions for multi-label image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5513–5522.
89. Yaghoubi, E.; Borza, D.; Neves, J.; Kumar, A.; Proença, H. An Attention-Based Deep Learning Model for Multiple Pedestrian Attributes Recognition. *arXiv preprint arXiv:2004.01110* **2020**.

90. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
91. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. European conference on computer vision. Springer, 2014, pp. 391–405.
92. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* **2015**.
93. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
94. Goodfellow, Ian J., P.A.J.M.M.X.B.W.F.D.O.S.C.A.C.; Bengio, Y. Generative adversarial nets. *NIPS* **2014**.
95. Kim, B.; Shin, S.; Jung, H. Variational autoencoder-based multiple image captioning using a caption attention map. *Applied Sciences* **2019**, *9*, 2699.
96. Xu, W.; Keshmiri, S.; Wang, G. Adversarially approximated autoencoder for image generation and manipulation. *IEEE Transactions on Multimedia* **2019**, *21*, 2387–2396.
97. Bekele, E.; Lawson, W.E.; Horne, Z.; Khemlani, S. Implementing a robust explanatory bias in a person re-identification network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2165–2172.
98. Yiru Zhao, Xu Shen, Z.J.H.L.X.s.H. Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification. Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4913–4922.
99. Chiat-Pin Tay, Sharmili Roy, K.H.Y. AANet: Attribute Attention Network for Person Re-Identifications. Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7134–7143.
100. Raza, M.; Zonghai, C.; Rehman, S.; Zhenhua, G.; Jikai, W.; Peng, B. Part-Wise Pedestrian Gender Recognition Via Deep Convolutional Neural Networks. 2nd IET International Conference on Biomedical Image and Signal Processing (ICBISP 2017). Institution of Engineering and Technology, 2017. doi:10.1049/cp.2017.0102.
101. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. AAAI, 2020, pp. 13001–13008.
102. Yaghoubi, E.; Borza, D.; Alirezazadeh, P.; Kumar, A.; Proença, H. Person Re-identification: Implicitly Defining the Receptive Fields of Deep Learning Classification Frameworks. *arXiv preprint arXiv:2001.11267* **2020**.
103. Chen, Q.; Huang, J.; Feris, R.; Brown, L.M.; Dong, J.; Yan, S. Deep domain adaptation for describing people based on fine-grained clothing attributes. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5315–5324.
104. Li, D.; Zhang, Z.; Chen, X.; Huang, K. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing* **2018**, *28*, 1575–1590.
105. Martinho-Corbishley, D.; Nixon, M.S.; Carter, J.N. Soft biometric retrieval to describe and identify surveillance images. 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). IEEE, 2016, pp. 1–6.
106. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. CBAM: Convolutional Block Attention Module. The European Conference on Computer Vision (ECCV), 2018.
107. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems, 2016, pp. 4898–4906.
108. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**.
109. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. VRSTC: Occlusion-Free Video Person Re-Identification. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
110. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
111. Hui Han, Wen-Yuan Wang, B.H.M. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *International Conference on Intelligent Computing, Springer* **2015**.

112. Haibo He, E.A.G. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **2009**.
113. Haibo He, Yang Bai, E.A.G.S.L. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks* **2008**.
114. Yuchun Tang, Yan-Qing Zhang, N.V.C.; Krasser, S. SVMs Modeling for Highly Imbalanced Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2008**.
115. Liu, Z.H.Z..X.Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* **2005**.
116. Abe, B.Z..J.L..N. Cost-sensitive learning by cost-proportionate example weighting. *Third IEEE International Conference on Data Mining* **2003**.
117. Nitesh V Chawla, Kevin W Bowyer, L.O.H.W.P.K. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research,* **2002**.
118. Hui Han, Wen-Yuan Wang, B.H.M. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International Conference on Intelligent Computing* **2005**.
119. Jo, T.; Japkowicz., N. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter* **2004**.
120. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* **2017**, *73*, 220–239.
121. GMiroslav Kubat, Stan Matwin, e.a. Addressing the curse of imbalanced training sets: one-sided selection. *ICML* **1997**.
122. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
123. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.
124. Yamaguchi, K.; Okatani, T.; Sudo, K.; Murasaki, K.; Taniguchi, Y. Mix and Match: Joint Model for Clothing and Attribute Recognition. BMVC, 2015, Vol. 1, p. 4.
125. Yamaguchi, K.; Berg, T.L.; Ortiz, L.E. Chic or social: Visual popularity analysis in online fashion networks. Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 773–776.
126. Deng, Y.; Luo, P.; Loy, C.C.; Tang, X. Pedestrian attribute recognition at far distance. Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 789–792.
127. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.
128. Li, D.; Zhang, Z.; Chen, X.; Ling, H.; Huang, K. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054* **2016**.
129. Sudowe, P.; Spitzer, H.; Leibe, B. Person attribute recognition with a jointly-trained holistic cnn model. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 87–95.
130. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* **2009**, *32*, 1627–1645.
131. Hall, D.; Perona, P. Fine-grained classification of pedestrians in video: Benchmark and state of the art. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5482–5491.
132. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision* **2010**, *88*, 303–338.
133. Xiong, Y.; Zhu, K.; Lin, D.; Tang, X. Recognize complex events from static images by fusing deep channels. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1600–1609.
134. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking, 2016.
135. Zhu, J.; Liao, S.; Lei, Z.; Yi, D.; Li, S. Pedestrian attribute classification in surveillance: Database and evaluation. Proceedings of the IEEE international conference on computer vision workshops, 2013, pp. 331–338.
136. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* **2013**.

137. Bileschi, S.M.; Wolf, L. CBCL streetscenes. Technical report, Center for Biological and Computational Learning (CBCL) at MIT, 2006.
138. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, Vol. 1, pp. 886–893.
139. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. Proceedings of the IEEE International Conference on Computer Vision, 2015.
140. Aruna Kumar, S.; Yaghoubi, E.; Das, A.; Harish, B.; Proença, H. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, Re-Identification and Search from Aerial Devices. *arXiv* **2020**, pp. arXiv–2004.
141. Chen, X.; Pang, A.; Zhu, Y.; Li, Y.; Luo, X.; Zhang, G.; Wang, P.; Zhang, Y.; Li, S.; Yu, J. Towards 3D Human Shape Recovery Under Clothing. *CoRR* **2019**, *abs/1904.02601*, [1904.02601].
142. Bertiche, H.; Madadi, M.; Escalera, S. CLOTH3D: Clothed 3D Humans. *arXiv preprint arXiv:1912.02792* **2019**.
143. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437* **2018**.
144. Barekatain, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Okutama-action: An aerial view video dataset for concurrent human action detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28–35.
145. Perera, A.G.; Law, Y.W.; Chahl, J. Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition. *Drones* **2019**, *3*, 82.
146. Zhang, S.; Zhang, Q.; Yang, Y.; Wei, X.; Wang, P.; Jiao, B.; Zhang, Y. Person Re-identification in Aerial imagery. *IEEE Transactions on Multimedia* **2020**, p. 1–1. doi:10.1109/tmm.2020.2977528.
147. Zhu, S.; Fidler, S.; Urtasun, R.; Lin, D.; Loy, C.C. Be Your Own Prada: Fashion Synthesis with Structural Coherence. International Conference on Computer Vision (ICCV), 2017.
148. Phillips, P.J.; Jiang, F.; Narvekar, A.; Ayyad, J.; O'Toole, A.J. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)* **2011**, *8*, 1–11.
149. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. Field and service robotics. Springer, 2018, pp. 621–635.