

Curriculum Learning of Visual Attribute Clusters for Multi-Task Classification

Nikolaos Sarafianos^a, Theodore Giannakopoulos^b, Christophoros Nikou^c, Ioannis A. Kakadiaris^a

^a*Computational Biomedicine Lab, Department of Computer Science
University of Houston, 4800 Calhoun Rd. Houston, TX 77004*

^b*National Center of Scientific Research Demokritos, Athens GR 15310, Greece*

^c*Department of Computer Science and Engineering, University of Ioannina, Ioannina GR 45110, Greece*

Abstract

Visual attributes, from simple objects (e.g., backpacks, hats) to soft-biometrics (e.g., gender, height, clothing) have proven to be a powerful representational approach for many applications such as image description and human identification. In this paper, we introduce a novel method to combine the advantages of both multi-task and curriculum learning in a visual attribute classification framework. Individual tasks are grouped after performing hierarchical clustering based on their correlation. The clusters of tasks are learned in a curriculum learning setup by transferring knowledge between clusters. The learning process within each cluster is performed in a multi-task classification setup. By leveraging the acquired knowledge, we speed-up the process and improve performance. We demonstrate the effectiveness of our method via ablation studies and a detailed analysis of the covariates, on a variety of publicly available datasets of humans standing with their full-body visible. Extensive experimentation has proven that the proposed approach boosts the performance by 4% to 10%.

Keywords: Curriculum Learning, Multi-task classification, Visual Attributes

1. Introduction

*Vision as reception. Vision as reflection. Vision as projection.
–Bill Viola, note 1986*

When we are interested in providing a description of an object or a human, we tend to use visual attributes to accomplish this task. For example, a laptop can have a wide screen, a silver color, and a brand logo, whereas a human can be tall, female, wearing a blue t-shirt and carrying a backpack. Visual attributes in computer vision are equivalent to the adjectives in our speech. We rely on visual attributes since (i) they enhance our understanding by creating an image in our head of what this object or human looks like; (ii) they narrow down the possible related results when we want to search for a product online or when we want to provide a suspect description; (iii) they can be composed in different ways to create descriptions; (iv) they generalize well as with some fine-tuning they can be applied to recognize objects for different tasks; and (v) they are a meaningful

semantic representation of objects or humans that can be understood by both computers and humans. However, effectively predicting the corresponding visual attributes of a human given an image remains a challenging task. In real-life scenarios, images might be of low-resolution, humans might be partially occluded in cluttered scenes, or there might be significant pose variations.

In this work, we introduce CILICIA (Curriculum Learning multiTask Classification Attributes) to address the problem of visual attribute classification from images of standing humans. Instead of using low-level representations, which would require extracting hand-crafted features, we propose a deep learning method to solve multiple binary classification tasks. CILICIA differentiates itself from the literature as: (i) it performs end-to-end learning by feeding a single ConvNet with the entire image of a human without making any assumptions about predefined connection between body parts and image regions; and (ii) it exploits the advantages of both multi-task and curriculum learning. Tasks are split into groups based on their labels' cross-correlation using hierarchical agglomerative clustering.

Email address: nsarafian@central.uh.edu (Nikolaos Sarafianos)

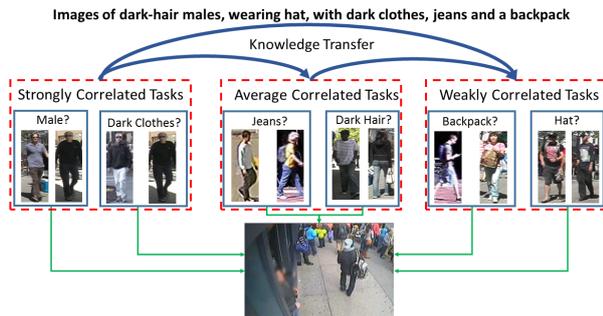


Figure 1: Curriculum learning for multi-task classification of visual attributes. Tasks are split into groups by performing hierarchical clustering which are then learned sequentially based on the cross-correlation of the attributes within each group. Flickr photo by Jeffery Scism is licensed under [CC BY](#).

The groups of tasks are learned in a curriculum learning scenario, starting with the one with the highest within-group cross-correlation and moving to the less correlated ones by transferring knowledge from the former to the latter. The tasks in each group are learned in a typical multi-task classification setup. Parts of this publication appear in our previous work [1]. However, in this work we have proposed an effective method to obtain the groups of tasks using hierarchical agglomerative clustering, which can be of any number and not just two groups (strongly/weakly correlated), conducted additional experiments to analyze the covariates of the proposed approach, benchmarked our method in an additional dataset, and demonstrated the efficacy and robustness of our method by performing ablation studies in Section 5.

When Vapnik and Vashist introduced the learning using privileged information (LUPI) paradigm [2], they drew inspiration from human learning. They observed how significant the role of an intelligent teacher was in the learning process of a student, and proposed a machine learning framework to imitate this process. Employing privileged information from an intelligent teacher at training time has recently received significant attention from the scientific community with remarkable results in areas ranging from object recognition [3, 4, 5, 6] to biometrics [7, 8, 9].

Our work also draws inspiration from the way students learn in class. First, students find it difficult to learn all tasks at once. It is usually easier for them to acquire some basic knowledge first, and then build on top of that, by learning more complicated concepts. This can be achieved by learning in a hierarchical manner, which is commonly employed in the literature

[10, 11, 12], or with a curriculum strategy. Curriculum learning [13, 14, 15] (presenting easier examples before more complicated and learning tasks sequentially, instead of all at the same time) imitates this learning process. It has the advantage of exploiting prior knowledge to improve subsequent classification tasks but it cannot scale up to many tasks since each subsequent task has to be learned individually. However, to maximize students’ understanding a curriculum might not be sufficient by itself. Students also need a teaching paradigm that can guide their learning process, especially when the task to be learned is challenging. The teaching paradigm in our method is the split of visual attribute classification tasks that need to be learned by performing hierarchical agglomerative clustering. In that way, we exploit the advantages of both multi-task and curriculum learning. First, the ConvNet learns the group of tasks with the strongest intra cross-correlation in a multi-task learning setup, and once this process is completed, the weights of the respective tasks are used as an initialization for the more diverse tasks. During the training of the more diverse tasks, the prior knowledge obtained is leveraged to improve the classification performance. An illustrative example of our method is depicted in Figure 1.

In summary, this paper has the following contributions. First, we introduce CILICIA, a novel method of exploiting the advantages of both multi-task and curriculum learning by splitting tasks into groups by performing hierarchical agglomerative clustering. The tasks of each subgroup are learned in a joint manner. Thus, the proposed method learns better than learning all the tasks in a typical multi-task learning setup and converges faster than learning tasks one at a time. Second, we propose a scheme of transferring knowledge between the groups of tasks which speeds up the convergence and increases the performance. We performed extensive evaluations in three datasets of humans standing and achieved state-of-the-art results in all three of them.

The remainder of the paper is organized as follows: in Section 2, a review of the related work in visual attributes, curriculum learning, and transfer learning is presented. Section 3 presents CILICIA, the proposed curriculum learning approach for multi-task classification of clusters of visual attributes. In Section 4, experimental results are reported, a detailed analysis of covariates is provided, and a discussion about the performance and the limitations of the proposed approach is offered. Finally, conclusions are drawn in Section 6.

2. Related Work

Visual Attributes Classification: The first to investigate the power of visual attributes were Ferrari and Zisserman [16]. They used low-level features and a probabilistic generative model to learn attributes of different types (e.g., appearance, shape, patterns) and segment them in an image. Kumar *et al.* [17] proposed an automatic method to perform face verification and image search. They first extracted and compared “high-level” visual features, or traits, of a face image that are insensitive to pose, illumination, expression, and other imaging conditions, and then trained classifiers for describable facial visual attributes (e.g., gender, race, and eyewear). A verification classifier on these outputs is finally trained to perform face verification. In the work of Scheirer *et al.* [18], raw attribute scores are calibrated to a multi-attribute space where each normalized value approximates the probability of that attribute appearing in the input image. This normalized multi-attribute space allows a uniform interpretation of the attributes to perform tasks such as face retrieval or attribute-based similarity search. Finally, attribute selection approaches have been introduced [19, 20, 21] which select attributes based on specific criteria (e.g., entropy). Zheng *et al.* [21] formulated attribute selection as a submodular optimization problem [22] and defined a novel submodular objective function.

Following the deep learning renaissance in 2012, several papers [23, 24, 25] have addressed the visual attribute classification problem using ConvNets. Part-based methods decompose the image to parts and train separate networks which are then combined at a feature level before the classification step. They tend to perform well since they take advantage of spatial information (e.g., patches that correspond to the upper body can better predict the t-shirt color than others that correspond to other body parts). Zhang *et al.* [26] proposed an attribute classification method which combines part-based models in the form of poselets [27], and deep learning by training pose-normalized ConvNets. Gkioxari *et al.* [28] proposed a deep version of poselets to detect human body parts which were then employed to perform action and attribute classification. Zhu *et al.* [29] introduced a method for pedestrian attribute classification. They proposed a ConvNet architecture comprising 15 separate subnetworks (i.e., one for each task) which are fed with images of different body parts to learn jointly the visual attributes. However, their method assumes that there is a pre-defined connection between parts and attributes and that all tasks depend on each other and thus, learn-

ing them jointly will be beneficial. Additionally, they trained the whole ConvNet end-to-end despite the fact that the size of the training dataset used was only 632 images. Based on our experiments, the only way to avoid heavy overfitting in datasets of that size is by employing a pre-trained network along with fine-tuning of some layers. Recycling pre-trained deep learning models with transfer learning (i.e., exploiting the discriminative power of a network trained for a specific task for a different problem or domain) is commonly used in the literature with great success [30, 31, 32]. Finally, visual attributes have been employed recently for re-identification [33, 34], pose estimation [35, 36], 3D pose tracking [37], head pose estimation and landmark detection [38], attribute mining and retrieval for clothing applications [39, 40, 41], zero-shot visual object categorization and recognition [42, 43], image annotation and segmentation [44], as well as surveillance [45].

Curriculum Learning: Solving all tasks jointly is commonly employed in the literature [46, 47, 48] as it is fast, easy to scale, and achieves good generalization. For an overview of deep multi-task learning techniques the interested reader is encouraged to refer to the work of Ruder [49]. However, some tasks are easier than others and also not all tasks are equally related to each other [50]. Curriculum Learning was initially proposed by Bengio *et al.* [13]. They argued that instead of employing samples at random it is better to present samples organized in a meaningful way so that less complex examples are presented first. Pentina *et al.* [50] introduced a curriculum learning-based approach to process multiple tasks in a sequence and developed a method to find the best order in which the tasks need to be learned. They proposed a data-dependent solution by introducing an upper-bound of the average expected error and employing an Adaptive SVM [51, 52]. Such a learning process has the advantage of exploiting prior knowledge to improve subsequent classification tasks but it cannot scale up to many tasks since each subsequent task has to be learned individually. Curriculum learning has also been employed with great success on long short-term memory (LSTM) networks [53, 54]. In parallel with our work, Dong *et al.* [55] also proposed a multi-task curriculum transfer technique to classify clothes based on their attributes. To the best of our knowledge, the method of Dong *et al.* [55] and our method, are the first to attempt formulating a curriculum learning strategy for a deep attribute classification problem.

Transfer Learning: Deep transfer learning techniques learn feature representations, which are transferable to other domains, by incorporating the adaptation to a new domain in the end-to-end learning process [56, 57]. The

idea of distilling knowledge in neural networks was initially introduced by Hinton *et al.* [58]. The authors proposed a method to distill the knowledge of a complex ensemble of models into a smaller model. The softmax output of the ensemble is divided by a temperature parameter and the smaller model learns directly from that “softened” output. Following that idea, Zhang *et al.* [59] suggested a technique to perform action recognition in real-time. They transferred knowledge from the teacher (an optical flow ConvNet) to the student (a motion vector ConvNet) by backpropagating the teacher’s loss in the students’ network. Finally, Lopez-Paz *et al.* [60] introduced generalized distillation; a method that unifies the LUPI framework with the knowledge distillation paradigm.

Finally, a very interesting prior work which focuses on the correlation of visual attributes is the method of Jayaraman *et al.* [61]. Aiming to decorrelate attributes at learning time, the authors proposed a multi-task learning framework with the property of resisting the urge of sharing image features of correlated attributes. Their approach disambiguates attributes by isolating distinct low-level features for distinct properties (e.g., color for “brown”, texture for “furry”). They also leveraged side information for properties that are closely related and should share features (e.g., “brown” and “red” are likely to share the same features). While our work also leverages information from correlated attributes in a multi-task classification framework, it models co-occurrence between different clusters of visual attributes instead of trying to semantically decorrelate them.

3. Methodology

In this section, we describe the proposed network architecture which given images of humans as an input, outputs visual attribute predictions. We then introduce our approach for splitting attributes into clusters. Finally, the proposed multi-task curriculum learning framework is introduced.

In our supervised learning paradigm, we are given tuples (x_i, y_i) where x_i corresponds to images and y_i to the respective visual attribute labels. The total number of tasks will be denoted by T , and thus the size of y_i for one image will be $1 \times T$. Finally, we will refer to the part of the network that solves the i^{th} group of tasks as C_i .

3.1. Multi-label ConvNet architecture

To mitigate the lack of training data we employ the pre-trained VGG-16 [62] network. VGG-16, is the net-

work from Simonyan and Zisserman which was one of the first methods to demonstrate that the depth of the network is a critical component for good performance. We selected VGG instead of a more modern network for the reason that it is a simple and homogeneous architecture, which despite its inefficiencies (e.g., large number of parameters), is sufficient for solving multiple binary image classification tasks. VGG-16 is trained on ImageNet [63], the scale of which enables us to perform transfer learning between ImageNet and our tasks of interest. The architecture of the network we use is depicted in Figure 2. We used the first seven convolutional layers of the VGG-16 network and dropped the rest of the convolutional and fully-connected layers. The reason behind this is that the representations learned in the last layers of the network are very task dependent [32] and thus, not transferable. Following that, for every task we added a batch-normalized [64] fully-connected layer with 512 units and a ReLU activation function. We employed batch-normalization since it enabled higher learning rates, faster convergence, and reduced overfitting. Although shuffling and normalizing each batch has proven to reduce the need of dropout, we observed that adding a dropout layer [65] was beneficial as it further reduced overfitting. The dropout probability was 75% for datasets with less than 1,000 training samples and 50% for the rest. For every task, an output layer is added with a softmax activation function using the categorical cross entropy.

Furthermore, we observed that the random initialization of the parameters of the last two layers back-propagated large errors in the whole network even if we used different learning rates throughout our network. To address this behavior of the network, which is thoroughly discussed in the method of Sutskever *et al.* [66], we “freeze” the weights of the pre-trained part and train only the last two layers for each task in order to learn the layer weights and the parameters of the batch-normalization.

After we ensured that we can always overfit on the training set, which means that our network is deep enough and discriminative enough for the tasks of interest, our primary goal was to reduce overfitting. Towards this direction, we (i) selected 512 units for the fully connected layer to prevent the network from learning several weights; (ii) employed a small weight decay of 10^{-4} for the layers that are trained; (iii) initialized the learning rate at 10^{-3} and reduced it by a factor of 5 every 100 epochs and up to five times in total; and (iv) augmented the data by performing random scaling up to 150% of the initial image followed by random crops, horizontal flips and adding noise by applying PCA to the RGB

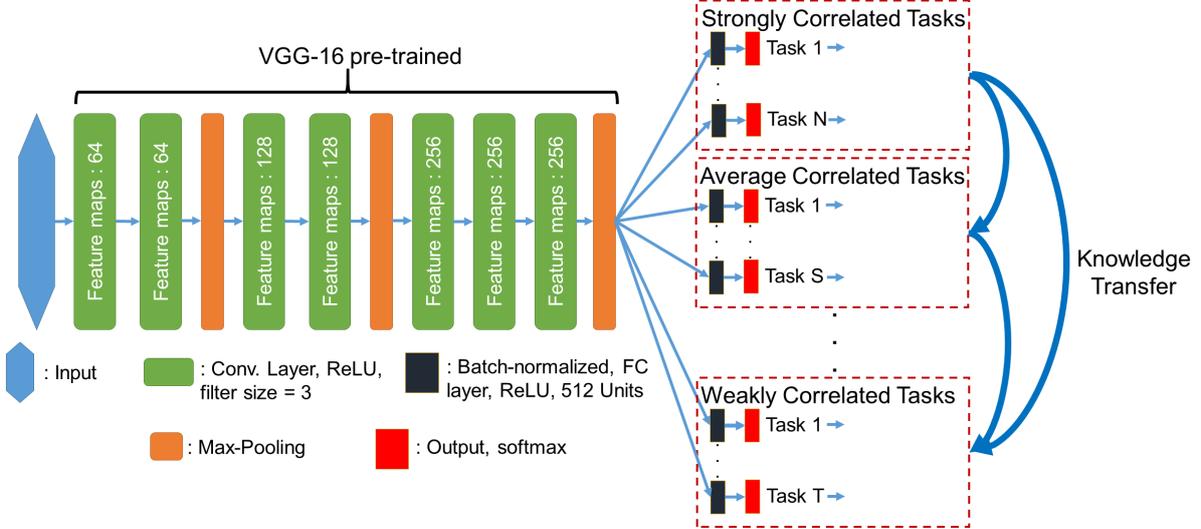


Figure 2: Architecture of the ConvNet used in our framework for several groups of tasks. The VGG-16 pre-trained part is kept frozen during training and only the weights of the last layers are learned. The different groups of tasks are learned sequentially using a curriculum learning paradigm. However, when the latter groups of tasks are trained, the tasks which have already been learned, contribute to the total cost function (Figure best viewed in color.)

pixel values as proposed by Krizhevsky *et al.* [67]. At test time, we averaged the predictions at three different scales (100%, 125%, and 150%) of five fixed crops and their horizontal flips (30 in total) to obtain the predicted class label. This technique, which was also adopted in the ResNet method of He *et al.* [68], proved to be very effective as it reduced the variation on the predictions.

3.2. Group Split with Hierarchical Clustering

Finding the order in which tasks need to be learned so as to achieve the best performance is difficult and computationally expensive. Given some tasks $t_i, i = 1 \dots T$ that need to be performed, we seek to find the best order in which the tasks should be performed so the average error of the tasks is minimized:

$$\text{minimize}_{S(t_i)} \frac{1}{T} \sum_{j=1}^T \mathcal{E}(\hat{y}_{t_j}, y_{t_j}), \quad (1)$$

where $S(t_i)$ is the function that finds the sequence of the tasks, \hat{y}_{t_j}, y_{t_j} are the prediction and target vectors for the j^{th} task, and \mathcal{E} is the prediction error.

However, the fact that a task can be easily performed does not imply that it is positively correlated with another and that by transferring knowledge the performance of the latter will increase. Adjeroh *et al.* [70] studied the correlation between various anthropometric

features and demonstrated that some correlation clusters can be derived in human metrology, whereby measurements in a cluster tend to be highly correlated with each other but not with the measurements in other clusters. The correlation between different sub-problems was also exploited in the age estimation method of Niu *et al.* [71] in an ordinal regression setup.

In this work, we seek to find: (i) which tasks (i.e., attributes) should be grouped together so as to be learned jointly, and (ii) which is the best sequence in which the groups of tasks should be learned. We use the training labels Y of size $N \times M$ where N the number of samples, and M the number of attributes (i.e., ground truth labels) to compute the Pearson correlation coefficient matrix which is of size $M \times M$. Each element in this matrix, represents to what extent these two attributes are correlated (e.g., the “gender” with the “hair length” will have a higher value compared to “gender” with “age”).

We then employ the computed Pearson correlation coefficient matrix to perform hierarchical agglomerative clustering using the Ward variance minimization algorithm. Ward’s method is biased towards generating clusters of the same size and analyzes all possible pairs of joined clusters, identifying which joint produces the smallest within cluster sum of squared (WCSS) errors. Assume that at an intermediate step, clusters s and t are to be merged to form cluster $u = s \cup t$. Then, the new distance $d(u, v)$ between cluster u and an already exist-

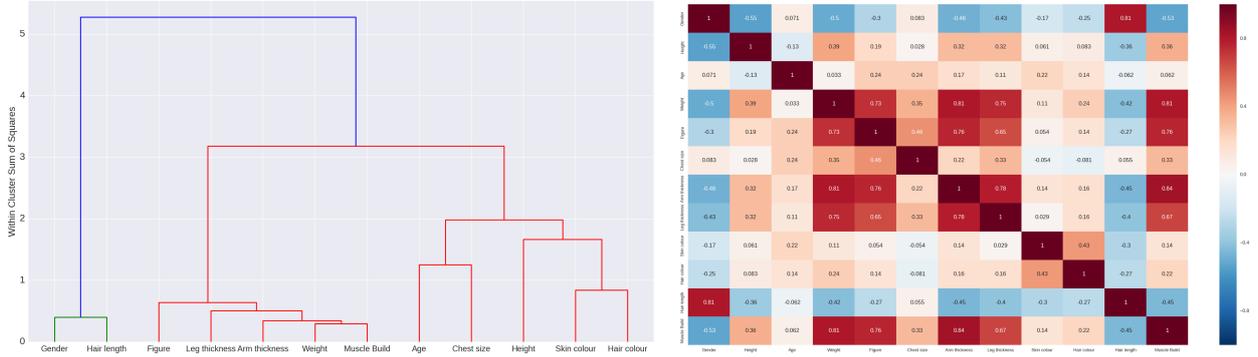


Figure 3: Dendrogram illustrating the arrangement of clusters (left) and the pairwise correlation matrix, which is fed to the clustering algorithm (right) of the visual attributes from the SoBiR dataset [69]. The sequence in which the clusters of attributes are learned, is obtained by computing the total dependency of each task with the rest within its cluster using Eq. (3). The curriculum learning of the clusters of visual attributes is then performed in a descending order as described in Section 3.3. (Figure best viewed in color.)

ing (but yet unused) cluster v is defined as:

$$d(u, v) = \sqrt{\frac{|v| + |s|}{T}d(v, s)^2 + \frac{|v| + |t|}{T}d(v, t)^2 + \frac{|v|}{T}d(s, t)^2}, \quad (2)$$

where s, t are the clusters which are joined into cluster u , and $T = |v| + |s| + |t|$. Ward [72], points out that this procedure facilitates the identification of that union which has an objective function value “equal or better than” any of the $n(n-1)/2$ possible unions. An illustrative hierarchical clustering of the visual attributes from the SoBiR dataset [69] in the form of a dendrogram is depicted in Figure 3. We observe that the proposed method for task split yields clusters of visual attributes which cohere with our semantic understanding and intuition about which attributes might be related to each other (e.g., gender with hair length, weight with muscle build). In addition to the pairwise correlation matrix, which also provides an insight into the relation of attributes, the proposed approach exploits this correlation between the attributes during the learning process.

By splitting the attributes into clusters using a WCSS threshold τ to cut the dendrogram horizontally, we have identified which tasks should be grouped together so as to be learned jointly. Following that, we now seek to obtain the sequence in which the clusters of visual attributes will be learned. To address this problem, we propose to find the total dependency $p_{i,c}$ of task $t_{i,c}$ with the rest within the cluster c , by computing the respective Pearson correlation coefficients but this time only within the cluster as follows:

$$p_{i,c} = \sum_{j=1, j \neq i}^T \frac{\text{cov}(y_{t_{i,c}}, y_{t_{j,c}})}{\sigma(y_{t_{i,c}})\sigma(y_{t_{j,c}})}, \quad i = 1, \dots, T \quad (3)$$

Algorithm 1: Finding the learning sequence of attribute clusters

- Input :** Training labels Y , WCSS threshold τ
- 1 $P \leftarrow$ compute Pearson correlation coefficient matrix split based on labels Y
 - 2 $G \leftarrow$ split into clusters using Eq. (2) along with P , labels Y , and τ
 - 3 **for** group g_i in G **do**
 - 4 $S_i \leftarrow$ compute average of cross-correlation within g_i using Eq. (3)
 - 5 **end**
 - 6 $S(g_i) \leftarrow$ compute learning sequence of clusters by sorting S_i 's in a descending order
- Output:** Learning sequence of clusters of visual attributes $S(g_i)$
-

where $\sigma(y_{t_{i,c}})$ is the standard deviation of the labels y of the task $t_{i,c}$. After we compute the total dependencies for all the clusters formed, we start the curriculum learning process in a descending order.

The process of computing the learning sequence of attribute clusters, which is described in detail in Algorithm. 1, is performed once before the training starts. Since it only requires the training labels of the tasks to compute the cross-correlations and perform the clustering, it is not computationally intensive. Finally, note that the group split depends on the training set and it is possible that different train-test splits might yield different groups of tasks.

Algorithm 2: Multi-task curriculum learning training

Input : Training set X , training labels Y , learning sequence of clusters $S(g_i)$ from Algorithm 1

- 1 **for** group g_i in $S(g_i)$ **do**
- 2 Initialize C_i from rest of already trained groups of tasks (if any)
- 3 $C_i \leftarrow$ train model using (X, Y_i) by minimizing the loss in Eq. (5)
- 4 **end**

Output: Parameters of network containing all groups of tasks

3.3. Multi-Task Curriculum Learning

In the scenario we are investigating, we solve multiple binary unbalanced classification tasks simultaneously. Throughout the paper, we use the terms multi-label and multi-task interchangeably. This is because we solve multiple classification tasks at the same time (multi-task), and at the same time for each given image we predict multiple binary labels which are not mutually exclusive (multi-label).

The proposed learning paradigm is described in Algorithm 2. Similar to Zhu *et al.* [73], we employ the categorical cross-entropy function between predictions and targets, which for a single attribute t is defined as follows:

$$L_t = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M \left(\frac{1/M_j}{\sum_{n=1}^M 1/M_n} \right) \cdot \mathbb{1}[y_i = j] \cdot \log(p_{i,j}), \quad (4)$$

where $\mathbb{1}[y_i = j]$ is equal to one when the ground truth of sample i belongs to class j , and zero otherwise, $p_{i,j}$ is the respective prediction, which is the output of the softmax nonlinearity of sample i for class j , and the term inside the parenthesis is a balancing parameter required due to imbalanced data. The total number of samples belonging to class j is denoted by M_j , N is the number of samples and M the number of classes. The total loss over all attributes is defined as $\sum_{t=1}^T \lambda_t \cdot L_t$, where λ_t is the contribution weight of each parameter. For simplicity, it is set to $\lambda_t = 1/T$. By setting λ_t in this way, there is an underlying assumption that all tasks contribute equally to the multi-task classification problem. To overcome this limitation, a fully-connected layer with T units could be added with an identity activation function after each separate loss L_t is computed. In that way, the respective weight for each attribute in the total loss function could be learned. However, we observed that for groups of

tasks that consist of a few attributes there was no difference in the performance, and thus we did not investigate this any further.

Once the classification of the visual-attribute tasks that demonstrated the strongest intra correlation is performed, we use the learned parameters (i.e., weights, biases, and batch normalization parameters) to initialize the network for the less diverse groups of attributes. The architecture of the network remains the same, with the parameters of VGG-16 being kept “frozen”. The weights of the tasks of previous groups of clusters continue to be learned with a very small learning rate of 10^{-6} . Furthermore, by adopting the “supervision transfer” technique of Zhang *et al.* [59] we leverage the knowledge learned by backpropagating the following loss:

$$L_j = \lambda \cdot L_t + (1 - \lambda) \cdot L_j^f, \quad (5)$$

where L_j^f is the total loss computed during the forward pass using Eq. (4) over only the current group of correlated tasks and λ is a parameter that controls the amount of knowledge transferred. Since the parameters of the network that correspond to already trained groups of tasks keep being updated, the loss L_t changes during training of the tasks of interest each time. This enables us to transfer the knowledge from groups of tasks with stronger intra cross-correlation to groups which demonstrated less intra cross-correlation. This technique proved to be very effective, as it enhanced the performance of the parts of the network which are responsible for the prediction of less correlated groups of tasks, and contributed to faster convergence during training.

4. Experiments

4.1. Datasets

To verify the effectiveness of the proposed method, we conducted evaluations in three challenging datasets containing standing humans, and thus tested our method in almost all the possible variations that can be found in the datasets used in the literature. We used the SoBiR [69], VIPeR [74] and PETA [75] datasets. The selected datasets are of varying difficulty and contain different visual attributes and training set sizes. Some representative images are depicted in Figure 4.

SoBiR dataset: The recently introduced SoBiR dataset [69] contains 800 images of 100 people. It comprises 12 soft biometric labels (e.g., gender, weight, age, height) and four forms of comprehensive human annotation. In our experimental investigation, we used the comparative ground-truth annotations (e.g., taller/shorter instead

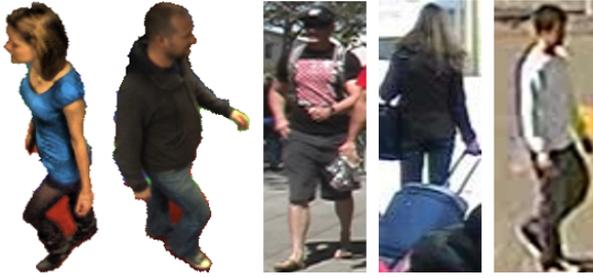


Figure 4: Example images from SoBiR, VIPeR and PETA datasets.

of tall/short) instead of absolute binary. The main reasons for this choice are: (i) relative binary annotations have been shown to outperform categorical annotations [69, 76]; and (ii) class labels were balanced for all soft biometrics. A 80/10/10 train/validation/test split based on human IDs is performed (so that only new subjects appear at testing) and average classification results are reported over five random splits.

VIPeR dataset: The VIPeR dataset [74] contains 632 low-resolution pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions. Layne *et al.* [77] provided 21 visual-attribute annotations which are used in our evaluation. We randomly split VIPeR into non-overlapping training and testing sets of equal sizes based on the human IDs. Following the literature, we repeated this process six times (one split for parameter tuning and the rest for evaluation of our method) and average classification results are reported.

PETA dataset: The PETA dataset [75] consists of 19,000 images gathered from 10 different smaller datasets. Parameters such as the camera angle, viewpoint, illumination, and resolution are highly variant, which makes it a valuable dataset for visual-attribute classification evaluation. It is divided in 9,500, 1,900, and 7,600 images for training, validation, and testing, respectively. Similar to [73], highly imbalanced attributes are discarded and the remaining 45 binary visual attributes are employed.

4.2. Results on SoBiR

Implementation details: For the SoBiR dataset, the batch size was set to 160. We split it into four clusters containing 2, 5, 2, and 3 attributes by thresholding at within cluster sum of squares $\tau = 1.9$ (Figure 3), trained our models for 5,000 epochs, and set $\lambda = 0.25$.

Evaluation results: Since the SoBiR dataset does not have a baseline on attribute classification we reported

results using handcrafted features and an SVM classifier as well as three different end-to-end learning frameworks using our ConvNet architecture. In all cases, images were resized to 128×128 . The features used for training the SVMs consisted of: (i) edge-based features, (ii) local binary patterns (LBPs), (iii) color histograms, and (iv) histograms of oriented gradients (HOGs). To preserve local information, we computed the aforementioned features in four blocks for every image resulting in 540 features in total. Furthermore, we investigated the classification performance when tasks are learned individually (i.e., by backpropagating only their own loss in the network), jointly in a typical multi-task classification setup (i.e., by backpropagating the average of the total loss in the network), and using the proposed approach. We report the classification accuracy (%) for all 12 soft biometrics in Table 1. CILICIA is superior in both groups of tasks to the rest of the learning frameworks. Despite the small size of the dataset, ConvNet-based methods perform better in all tasks compared to an SVM with handcrafted features. Multi-task learning methods (i.e., multi-task and CILICIA) outperform the learning frameworks when tasks are learned independently since they leverage information from other attributes. By taking advantage of the correlation between attributes, CILICIA demonstrated higher classification performance than a typical multi-task learning scenario. However, estimating the “age” proved to be the most challenging task in all cases as its classification accuracy ranges from 58.5% to 64.5% when it is learned individually using our ConvNet architecture. This poor performance can be attributed to the fact that age estimation from images without facial traits is a largely unsolved problem. In Figure 5, the convergence plots for all four CILICIA groups are depicted and the following observations are made: (i) the first group (comprises only two attributes) after epoch 3,000, demonstrates strong overfitting which proved to be inevitable even when we experimented with smaller learning rates; (ii) Multi-Task learning demonstrated the highest loss compared to the groups of the proposed method; and (iii) as we move from the groups of attributes that are strongly correlated to the rest by transferring knowledge each time, the training loss becomes smaller and there is less overfitting (if any). Note that the depicted losses for the corresponding groups are averaged over the tasks that belong to the cluster and thus, they can be compared although the number of tasks in each group is not the same.

Impact of color information: In this experiment, we investigated to what extent the color information affects the classification accuracy of the visual attributes. Since

Table 1: Classification accuracy of different learning paradigms on the SoBiR dataset. In individual learning, each attribute is learned separately. In multi-task learning, the average loss of all attributes is backpropagated in the network. In CILICIA, four clusters were formed and attributes are in descending order based on their intra cross-correlation. Results highlighted with light purple indicate statistically significant improvement using the paired-sample t-test.

Soft Label	SVM	Individual Learning	Multi-Task Learning	CILICIA
Gender	72.1	80.4	79.6	85.2
Height	64.7	73.9	72.0	77.0
Age	58.5	62.6	61.9	64.5
Weight	57.7	67.7	71.0	74.1
Figure	57.8	68.7	67.1	67.3
Chest size	58.7	64.9	68.9	67.5
Arm thickness	60.1	72.0	73.1	73.7
Leg thickness	56.7	68.9	71.0	72.6
Skin color	59.2	66.8	67.6	68.7
Hair color	67.5	74.2	76.1	77.9
Hair length	71.8	78.9	79.2	85.9
Muscle build	58.5	73.3	74.5	75.8
Total Av.	61.9	71.0	71.3	74.2

we employed a network pre-trained on RGB images, we fed each of the three color channels with the same grayscale images. Figure 6 (left) summarizes the obtained results. Visual attributes such as hair and skin color were expected to have a drop in performance when color information was absent. However, we observed that weight and age were also affected as their accuracy dropped by 8% and 11% respectively.

Impact of image resolution: The objective of this experiment is to assess the impact of image resolution on the classification performance. One of the major advantages of ConvNets (due to parameter sharing) is that they do not need to have a fixed-size input and pre-trained networks can be utilized with images of different spatial size. This still holds for fully-connected layers since they are equivalent to convolution layers with 1×1 kernel. Our ConvNet architecture was fed with images of varying resolutions starting from 32×32 up to 128×128 and average classification accuracies over all attributes are reported. From the obtained results in Figure 6 (right), we draw the following observations. First, images of higher resolution tend to perform better than lower since they provide more spatial space for the convolutional operations. Second, when we experimented with images of size 32×32 , we observed that the norm of the gradient started taking very high values which is a common phenomenon during training ConvNets and is

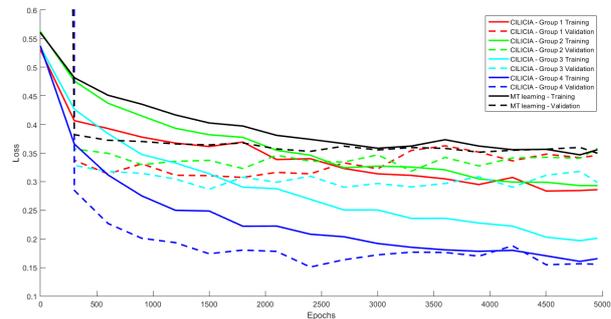


Figure 5: Convergence plot for all the groups of CILICIA as well as Multi-Task learning on the SoBiR dataset. Note that the numbering starts from the most correlated to the least correlated cluster. (Figure best viewed in color.)

referred to as the exploding gradients problem [78]. In our case, the reasons for the explosion of gradients were the small image size in a network pre-trained on images of almost 10-times larger and the existence of dropout with a high probability before the output layer.

4.3. Results on VIPeR

Implementation details: For the VIPeR dataset, the batch size was set to 158. We split it into five clusters containing 2, 4, 7, 3, and 4 attributes by thresholding at within cluster sum of squares $\tau = 1.8$, trained our models for 5,000 epochs, and set $\lambda = 0.25$.

Evaluation results: To demonstrate the superiority of the proposed approach over normal multi-task learning approaches, we evaluate in Table 2 its performance in comparison with the method of Zhu *et al.* [29] and a typical multi-task learning framework. To test for statistical significance between CILICIA and the method of Zhu *et al.* [29] we employed the z-test ($p < 0.05$) since the mean and the standard deviation results were available from their technique. Employing the proposed multi-task curriculum learning approach is beneficial for the classification of visual attributes, as it outperformed the previous state-of-the-art by improving the total results by 9.9%. CILICIA achieved significantly better results in most of the tasks, which demonstrates the efficacy of our method over traditional multi-task learning approaches. The reason for this is that when some tasks are completely unrelated then multi-task learning has a negative effect as it forces the network to learn representations that explain everything, which is not possible. Additionally, we observed that color attributes tend to achieve higher performance compared to other attributes. The reason for this is that such attributes are highly imbalanced (sometimes more than

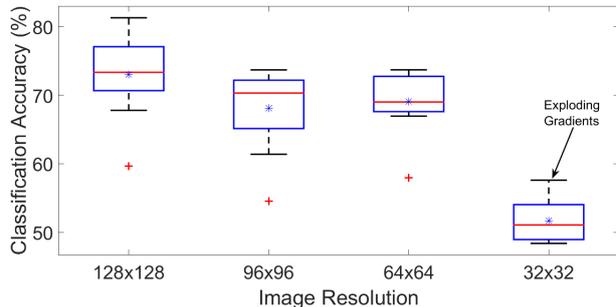
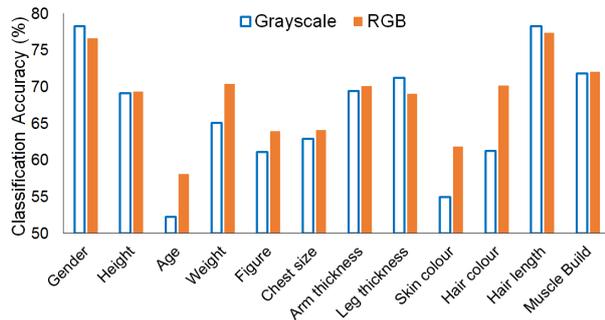


Figure 6: Impact of image color information in the classification accuracy of the attributes of the SoBiR dataset (left) and classification accuracy for different image resolutions (right). The standard deviation is high because we averaged over all attributes of the SoBiR dataset. Outliers depicted with a red cross correspond to “age”.

Table 2: Performance comparison on the VIPeR dataset. Five clusters were formed and attributes are in descending order based on their intra cross-correlation. Results highlighted with light purple indicate statistically significant improvement using the z-test.

Visual Attribute	Multi-Task Learning	Zhu <i>et al.</i> [29]	CILICIA
barelegs	79.6 ± 0.8	84.1 ± 1.1	88.6 ± 0.4
shorts	76.8 ± 1.1	81.7 ± 1.3	89.6 ± 0.6
lightshirt	79.5 ± 0.9	83.0 ± 1.2	84.6 ± 0.6
nocoats	74.3 ± 1.3	71.3 ± 0.8	75.0 ± 0.3
blueshirt	69.9 ± 1.7	69.1 ± 3.3	90.2 ± 1.2
midhair	74.3 ± 1.3	76.1 ± 1.8	77.7 ± 0.6
lightbottoms	79.0 ± 1.0	76.4 ± 1.2	75.1 ± 1.2
redshirt	79.2 ± 1.9	91.9 ± 1.0	93.6 ± 0.4
nolightdarkjeans	87.1 ± 1.6	90.7 ± 2.0	96.0 ± 0.5
greenshirt	70.3 ± 2.4	75.9 ± 5.9	95.1 ± 0.4
hashandbag	66.9 ± 3.1	42.0 ± 6.5	90.8 ± 0.6
hassatchel	72.5 ± 0.8	57.8 ± 2.7	72.8 ± 0.4
skirt	67.2 ± 3.7	78.1 ± 3.5	94.8 ± 0.6
darkbottoms	68.1 ± 0.9	78.4 ± 0.7	76.3 ± 0.9
male	71.5 ± 1.9	69.6 ± 2.6	81.0 ± 1.9
patterned	67.4 ± 3.5	57.9 ± 9.2	92.2 ± 0.7
darkshirt	71.0 ± 1.4	82.3 ± 1.4	84.3 ± 0.5
jeans	74.9 ± 0.7	77.5 ± 0.6	76.2 ± 0.6
darkhair	70.1 ± 2.0	73.1 ± 2.1	71.8 ± 1.3
hasbackpack	68.4 ± 1.4	64.9 ± 1.2	76.1 ± 1.6
Total Av.	73.4 ± 1.2	74.1 ± 1.0	84.0 ± 0.8

one to nine) due to the way annotation is provided (e.g., is the human wearing a red t-shirt or not). Note that, since in both the SoBiR and the VIPeR datasets, the training set is not fixed and train-test splits are performed each time, the sequence in which the groups of tasks are learned is not necessarily the same each time. The reason for this is that the hierarchical agglomerative clustering depends on the pairwise correlation matrix between the labels of the training set which vary between different splits.

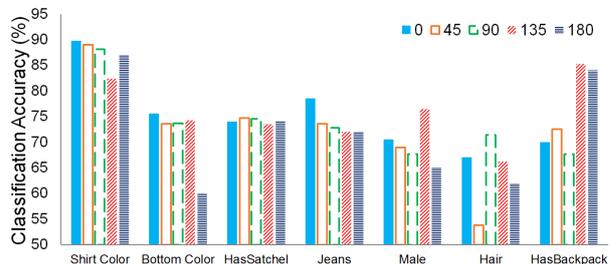


Figure 7: Classification performance for the low-correlated attributes under different viewpoints. For visualization purposes, we group attributes (e.g., “dark-hair” and “midhair” correspond to “hair”).

Impact of facial information: To investigate the impact of facial information in the classification, we trained the proposed architecture with images of humans after removing the upper part of the image containing the face (top 20% of the image). From the results in Table 3, we observed that the performance of the visual attributes related to clothes or objects of the upper body was not affected by the absence of facial information. On the contrary, the impact on the hair-related attributes was significant since their performance dropped by 6%. Finally, facial information plays a vital role in recognizing the gender of humans even in low-resolution images as a 5.3% performance drop was observed in the absence of the face.

Impact of viewpoint: Aiming to investigate to what extent the camera viewpoint affects the classification performance, we report classification results over the weakly-correlated attributes depending on the camera angle. VIPeR images are captured from 0°, 45°, 90°, 135°, and 180° degrees and the obtained results are depicted in Figure 7. It can be observed that visual attributes related to shirt color are viewpoint-

invariant, whereas others such as “jeans” perform better from a frontal angle. Two interesting observations arise from the hair and backpack attributes. First, classifying attributes pertaining to hair (length and color) can be done with higher accuracy when the viewpoint is at 90° . Second, finding whether the human has a backpack or not becomes an easier task for camera angles of 135° and 180° which is compatible with the way humans would perform on this task.

4.4. Results on PETA

Implementation details: For the PETA dataset, the batch size was set to 190. We split it into five clusters containing 2, 11, 4, 10, and 18 attributes by thresholding at within cluster sum of squares $\tau = 3$, trained our models for 5,000 epochs, and set $\lambda = 0.2$.

Evaluation results: Since the training size of the PETA dataset is significantly higher than the rest (almost 10,000) and the annotations provided are 45 instead of 20, some very interesting observations can be made from the clusters of visual attributes depicted in Figure 8. The turquoise cluster comprises attributes related to upper and lower body formal clothes along with black and leather footwear, and thus it is beneficial if we learn these attributes at the same time. Other examples that follow to our intuition and semantic understanding are the fact that being male is very strongly connected with having short hair and not carrying any type of bag, or that carrying a backpack is linked with being less than 30 years old. The proposed learning approach employs this information from attributes strongly connected on the PETA dataset and outperformed the recent method of Zhu *et al.* [73].

Since many attributes are highly imbalanced and the classification accuracy as an evaluation metric is not sufficient by itself they also reported recall rate results when the false positive rate is equal to 10% as well as the area under the ROC curve (AUC). Following the same evaluation protocol, we tested the proposed multi-task curriculum learning method on the PETA dataset and report our results in comparison with those of Zhu

Table 3: Performance comparison of the weakly correlated visual attributes when the face exists and it is absent in the input image. Attributes are grouped based on where they correspond to the human body.

	Upper Body	Lower Body	Hair	Gender
Entire Image	85.3	84.8	74.7	81.0
No Face	83.2	85.4	68.7	75.7

et al. [73] after grouping the attributes in Table 4. Although our method is not part-based, as it does not split the human image into parts which are then learned individually, it outperforms the part-based method of Zhu *et al.* [73] in all types of visual attributes under all evaluation metrics. Due to highly imbalanced data (the imbalance ratio in most of the categories is relatively high), the improvement in the classification accuracy is minor. However, for the rest of the evaluation metrics, our method improved the average recall rate by 3.93% and the AUC by 1.94%. In Figure 9 the ROC curves of some tasks in which our method performed really well (e.g., “blue shirt”), reasonably well (e.g., “gender”), and adequately (e.g., “has backpack”) are depicted. The complete results on the PETA dataset are provided in Table 6.

5. Ablation Studies and Performance Analysis

5.1. Is Hierarchical Clustering Beneficial?

An important question that arises while analyzing the performance of CILICIA on attribute classification is, what is the impact of the group split using hierarchical clustering along with the proposed learning paradigm which guides the training process? To what extent can the obtained results be accustomed to using a deep-learning multi-task classification scheme? To answer such questions, we conducted a detailed experimental evaluation on the VIPeR dataset [74] for a different number of groups that were split with two different methods. The first group split is performed after sorting the total cross-correlations for each attribute using Eq. (3), in descending order, and splitting the attributes to a fixed number of groups. For example, for the VIPeR dataset, we compute for each attribute the sum of cross-correlations with the rest, which results in a 20×1 vector. After sorting this vector, we split it into N number of groups depending on the number of groups that we are interested in investigating. The second way to split the visual attributes into groups is by employing the proposed approach (described in Section 3.2). In Figure 10, we demonstrate the average classification accuracy over all visual attributes of the VIPeR dataset [74] for a different number of groups, using the two methods described. We observe that splitting into groups using a hierarchical top-down method yields better results in all groups and that for the VIPeR dataset, five clusters are the optimal number of groups.

5.2. Why is Knowledge Transfer important?

To assess the impact of transferring knowledge from groups of tasks which have already converged to ones

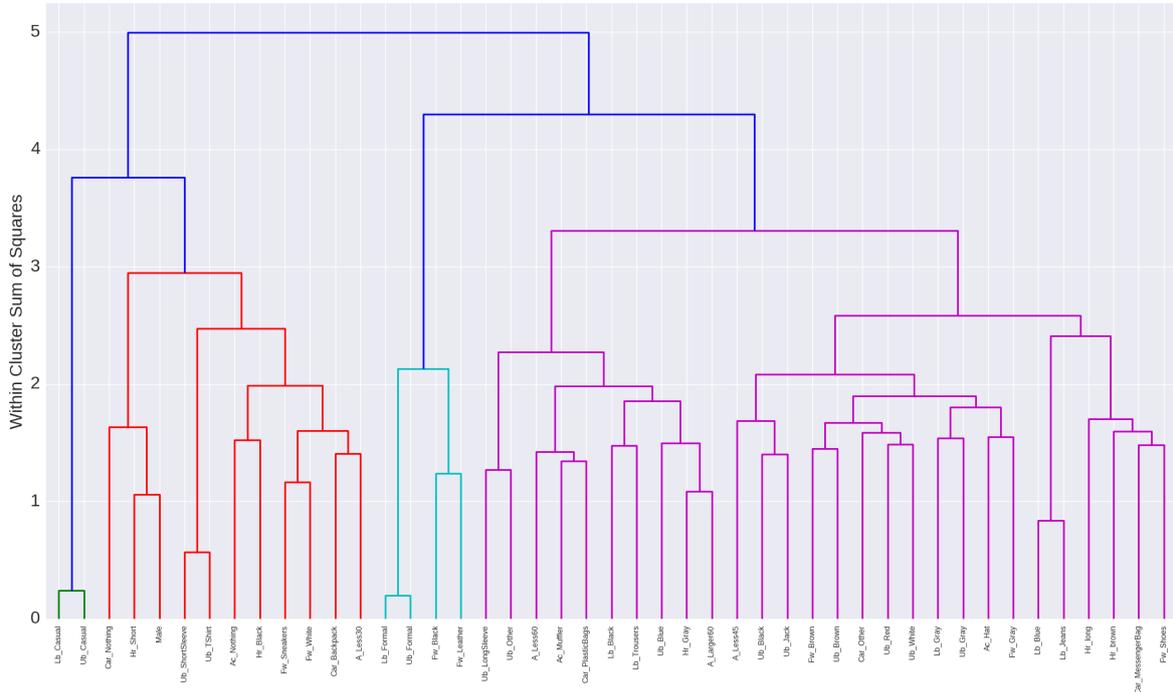


Figure 8: Dendrogram of the visual attributes of the PETA training set. For WCSS equal to 3, five clusters are formed. The learning sequence of the clusters is green, left purple, red, right purple and turquoise. (Lb. - Lower Body, Ub. - Upper Body, Car. - Carrying, Ac. - Accessory, Fw. Footwear, Hr. - Hair, and A. - Age).

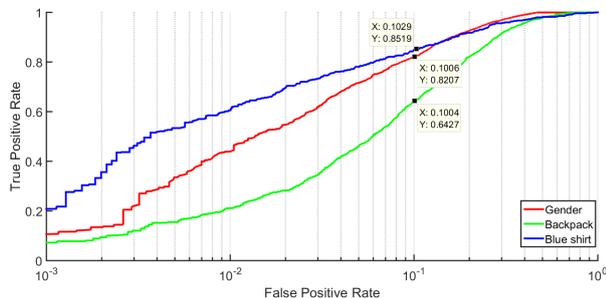


Figure 9: ROC curves for the visual attributes of “gender”, “blue shirt”, and “has backpack”. The x-axis is in semi-logarithmic scale and the depicted values correspond to the recall rate (%) when the false positive rate is 10%.

that have not been learned yet we conducted an ablation experiment. We selected the four most correlated and the four least correlated attributes of the PETA dataset so as to form the two groups of strongly and weakly correlated attributes. We compare the classification accuracy of the selected tasks with and without knowledge transfer. When no knowledge is transferred to the latter group, we are simply training two multi-task classi-

fication frameworks. We report the obtained results in the last two columns of Table 5. Transferring knowledge from a strongly correlated group of tasks to the weakly improves the performance of the latter by 1.89% compared to a typical multi-task classification learning framework.

5.3. Why use Correlation as a Criterion for Group Split?

To demonstrate the effectiveness of clustering attributes into groups based on their cross correlation we conducted an ablation study using the same eight attributes (as in Section 5.2) from the PETA dataset. However, in this experiment, instead of grouping them based on their cross-correlation, we randomly assign them to two groups. We follow exactly the same two-stage process (i.e., learning one group first and transferring knowledge to the second which is learned right after) and report the obtained results in the first column of Table 5. We observe that learning in correlation-based groups of tasks is beneficial as CILICIA with and without knowledge transfer performs better than learning at random. Additionally, transferring knowledge between attributes that do not co-occur (or they are semantically

Table 4: Performance comparison on the PETA dataset for different types of attributes. The imbalance ratio is defined as the ratio of the number of instances in the majority class to the number of examples in the minority class in the training set.

Visual Attribute	Imbalance Ratio	Accuracy (%)		Recall rate (%) @FPR = 10%		AUC (%)	
		Zhu <i>et al.</i> [73]	CILICIA	Zhu <i>et al.</i> [73]	CILICIA	Zhu <i>et al.</i> [73]	CILICIA
Accessories	7.63	93.11	93.48	75.68	76.66	91.06	92.13
Carrying Bags	5.01	83.68	84.78	57.21	62.73	82.79	85.63
Footwear	4.69	83.41	83.74	59.09	60.88	84.44	85.18
Hair	4.57	89.54	89.96	75.89	80.43	90.95	93.18
Lower Body	3.54	85.05	85.66	64.92	66.95	87.37	88.26
Upper Body	8.06	89.60	90.48	69.88	76.12	88.66	91.68
Age	7.05	87.84	87.90	71.03	72.49	88.93	90.24
Gender	1.22	84.34	87.59	74.80	82.04	91.74	93.84
Total Av.	6.07	87.23	87.91	67.29	71.22	87.66	89.60

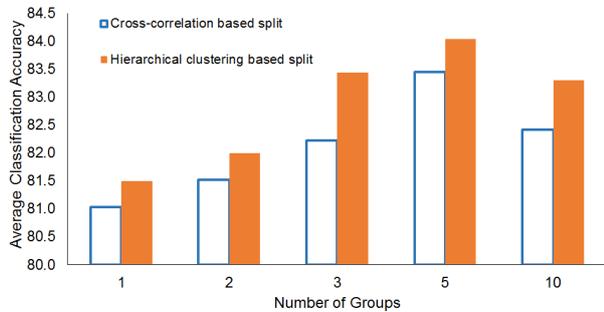


Figure 10: Average classification accuracy over all visual attributes of the VIPeR dataset, for different number of group splits. The cross-correlation based split refers to grouping the tasks based on their total cross-correlation with the rest after arranging them in a descending order. The hierarchical clustering based split corresponds to the proposed approach described in Algorithm 1.

completely different) has an adverse effect on the performance. The obtained results are in line with previous methods that can be found in the literature [79, 80] that have exploited label correlations to improve multi-task learning.

5.4. Why is the Proposed Curriculum the Right One?

We argue that task similarity and thus the curriculum is not binary, but resides on a spectrum. In the same way that humans learn with different curricula depending on the task, the process of finding a curriculum that is beneficial for all tasks cannot have an optimal single solution. Learning in correlation-split groups showed promising results (Tables 1 and 5) which led

Table 5: Ablation experiments to assess the effectiveness of knowledge transfer and correlation-based split. In the random split column, the strongly and weakly groups refer only to the learning sequence as the split is not based on the correlation. CILICIA (w/o kt) corresponds to learning in correlation-split groups but without knowledge transfer.

Group	Random Split	CILICIA (w/o kt)	CILICIA
Strongly	65.36	76.01	76.01
Weakly	63.08	69.91	71.80
Total	64.22	72.95	73.91

us to start considering how can we improve the performance. Transferring knowledge between related tasks is not beneficial as during the joint multi-task learning training the parameter sharing plays that role. Transferring knowledge from randomly-split groups also proved to be ineffective (Table 5). We then investigated whether the work of Bengio *et al.* [13], which proposed a curriculum based on what is easier to learn first, would add value. We believe that the knowledge transfer from the strongly to the weakly correlated group of tasks is a reasonable easy-to-hard curriculum which resembles to the definition of Bengio *et al.* [13]. In addition, note that when Bengio *et al.* [13] introduced curriculum learning after they defined an entropy-based curriculum they demonstrated that introducing gradually more difficult examples speeds-up online training. In our paper, this can be observed in the convergence plot (Figure 5) in which the subgroups converge faster and with a smaller loss (average among tasks).

5.5. Performance Analysis and Limitations

The proposed approach outperformed the state-of-the-art in all datasets and demonstrated better results over the rest of the learning paradigms on the SoBiR dataset. The main reasons for this are: (i) we exploited the correlation between different attributes and split them into clusters using hierarchical clustering; (ii) we proposed a learning paradigm to learn the group of attributes in a curriculum learning framework and classify them in a multi-task classification setup; and (iii) we leveraged the already learned clusters of visual attributes which had converged to transfer knowledge to groups that were about to be learned to improve the performance and enhance the stability of our method.

Despite its success and good performance, the proposed approach has a few limitations and inefficiencies. First, the existence of a fully-connected layer after the last convolutional layer increases significantly the number of parameters that need to be learned for each task. We partially addressed this by freezing most of the network and employed a small number of units in the fully-connected layer. This inefficiency is known for the VGG network and was addressed by more recent networks such as the GoogLeNet [81], the ResNet [68] or the Highway Networks [82]. However, these networks were not suitable for CILICIA because of the way training is performed (i.e., in all three networks, the same inputs travel through different paths and different number of layers). Second, the proposed approach contains two additional parameters that need to be cross-validated thoroughly. The first parameter is λ , which controls the contribution of the already learned groups of clusters and is found in several methods that perform transfer learning or knowledge distillation [60, 59]. For this parameter, we experimented on the validation set with different parameters (namely 0.25, 0.5, 0.75 and 1) and observed that a 25% contribution of the already learned clusters of visual attributes was the most effective. The second parameter is the within-cluster sum of squares threshold which controls the number of clusters formed. Finally, the goal of the proposed approach was to classify the visual attributes of humans, the full body of whom was always fully-visible. Thus, it was tested in re-identification datasets, which contain pairs of images of humans standing or walking, and outperformed the state-of-the-art without even following a part-based approach. For datasets such as the Berkeley Attributes of People dataset [27], which comprises humans of varying poses with parts of their body either not visible or occluded, part-based (or poselet-based) approaches [28, 24, 23] have proven to be very effective recently. To address these challenges, CILICIA would have to

be adapted to work with poselets or body-parts (i.e., in which order the body parts need to be learned so as to transfer information between groups of tasks) which was outside the scope of this paper.

6. Conclusion

Given a set of tasks that need to be learned we sought to find an answer to how we can learn them effectively and what would be the optimal way in terms of performance, speed, and simplicity. Learning each task separately, although very simple, lacks in terms of performance since it does not exploit the information from other tasks. Learning all tasks at the same time in a multi-task classification scenario is relatively fast, easy to implement, and employs knowledge from other tasks to boost the classification performance. Curriculum learning is a learning scheme in which samples or tasks are not treated as equally easy or hard, but are instead presented to the model in a meaningful way so as to increase generalization and performance. Since learning a large number of tasks one at a time is computationally expensive, we opted for learning clusters of tasks in a curriculum. In each cluster of visual attributes, we proposed to learn the corresponding tasks in a multi-task classification setup.

Our proposed method, CILICIA, finds the sequence in which clusters of visual attributes are learned very efficiently, and classifies them with high performance. Given images of standing humans as an input, we performed end-to-end learning by solving multiple binary classification problems simultaneously. Tasks were grouped into clusters by employing hierarchical agglomerative clustering based on their correlation. The sequence (i.e., curriculum) in which clusters were learned was found by computing the average cross-correlation within each cluster and sorting the obtained values in a descending order. During training of weakly correlated clusters of tasks, we leveraged the knowledge already learned from clusters which demonstrated stronger correlation. By these means, we combined the advantages of both multi-task and curriculum learning paradigms; since our method converges fast, it is effective and employs prior knowledge. We evaluated our method in three datasets and outperformed the state-of-the-art by 9.9% on the VIPeR dataset and by a recall rate of almost 4% (when the false positive rate is fixed and equal to 10%) on the PETA dataset despite the fact that no body part-specific information was employed. The obtained results demonstrate the effectiveness and, at the same time, the great potential of multi-task curriculum learning.

Acknowledgments

This work has been funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. The work of C. Nikou is supported by the European Commission (H2020-MSCA-IF-2014), under grant agreement No 656094. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

- [1] N. Sarafianos, T. Giannakopoulos, C. Nikou, I. A. Kakadiaris, Curriculum learning for multi-task classification of visual attributes, arXiv preprint arXiv:1708.08728. 2
- [2] V. Vapnik, A. Vashist, A new learning paradigm: Learning using privileged information, *Neural Networks* 22 (5-6) (2009) 544–57. 2
- [3] S. Motian, M. Piccirilli, D. A. Adjeroh, G. Doretto, Information bottleneck learning using privileged information for visual recognition, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016. 2
- [4] D. Pechyony, R. Izmailov, A. Vashist, V. Vapnik, SMO-style algorithms for learning using privileged information, in: *Proc. International Conference on Data Mining*, Las Vegas, NV, 2010. 2
- [5] V. Sharmanska, N. Quadrianto, C. H. Lampert, Learning to rank using privileged information, in: *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, 2013. 2
- [6] S. Wang, D. Tao, J. Yang, Relative attribute SVM+ learning for age estimation, *IEEE Transactions on Cybernetics* 46 (3) (2015) 827–839. 2
- [7] I. A. Kakadiaris, N. Sarafianos, C. Nikou, Show me your body: Gender classification from still images, in: *IEEE International Conference on Image Processing*, Phoenix, AZ, 2016. 2
- [8] M. Vrigkas, C. Nikou, I. A. Kakadiaris, Exploiting privileged information for facial expression recognition, in: *Proc. IEEE International Conference on Biometrics*, Halmstad, Sweden, 2016. 2
- [9] N. Sarafianos, C. Nikou, I. A. Kakadiaris, Predicting privileged information for height estimation, in: *Proc. International Conference on Pattern Recognition*, Cancun, Mexico, 2016. 2
- [10] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, HD-CNN: Hierarchical deep convolutional neural network for large scale visual recognition, in: *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, 2015. 2
- [11] L. Zhang, P. Dou, S. K. Shah, I. A. Kakadiaris, Hierarchical multi-label framework for robust face recognition, in: *Proc. IEEE International Conference on Biometrics*, Phuket, Thailand, 2015. 2
- [12] L. Zhang, S. K. Shah, I. A. Kakadiaris, Hierarchical multi-label classification using fully associative ensemble learning, *Pattern Recognition* 70 (2017) 89–103. 2
- [13] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: *Proc. International Conference on Machine Learning*, Montreal, Canada, 2009. 2, 3, 13
- [14] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, A. Hauptmann, Self-paced learning with diversity, in: *Advances in Neural Information Processing Systems*, Montreal, Canada, 2014. 2
- [15] T. Matisen, A. Oliver, T. Cohen, J. Schulman, Teacher-student curriculum learning, arXiv preprint arXiv:1707.00183. 2
- [16] V. Ferrari, A. Zisserman, Learning visual attributes, in: *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, 2007. 3
- [17] N. Kumar, A. Berg, P. N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (10) (2011) 1962–1977. 3
- [18] W. J. Scheirer, N. Kumar, P. N. Belhumeur, T. E. Boult, Multi-attribute spaces: Calibration for attribute fusion and similarity search, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012. 3
- [19] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009. 3
- [20] W. Wang, Y. Yan, S. Winkler, N. Sebe, Category specific dictionary learning for attribute specific feature selection, *IEEE Transactions on Image Processing* 25 (3) (2016) 1465–1478. 3
- [21] J. Zheng, Z. Jiang, R. Chellappa, Submodular attribute selection for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 3
- [22] A. Krause, D. Golovin, Submodular function maximization, *Tractability: Practical Approaches to Hard Problems* 3 (19) (2012) 1–8. 3
- [23] Y. Li, C. Huang, C. C. Loy, X. Tang, Human attribute recognition by deep hierarchical contexts, in: *Proc. European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016. 3, 14
- [24] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with R*CNN, in: *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, 2015. 3, 14
- [25] P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly-trained holistic CNN model, in: *Proc. IEEE International Conference on Computer Vision Workshops*, Boston, MA, 2015. 3
- [26] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, PANDA: Pose aligned networks for deep attribute modeling, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014. 3
- [27] L. Bourdev, S. Maji, J. Malik, Describing people: A poselet-based approach to attribute classification, in: *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011. 3, 14
- [28] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, 2015. 3, 14
- [29] J. Zhu, S. Liao, D. Yi, Z. Lei, S. Z. Li, Multi-label CNN based pedestrian attribute learning for soft biometrics, in: *Proc. IEEE International Conference on Biometrics*, Phuket, Thailand, 2015. 3, 9, 10
- [30] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, 2014. 3
- [31] J. Yang, B. Price, S. Cohen, H. Lee, M.-H. Yang, Object contour detection with a fully convolutional encoder-decoder network, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016. 3
- [32] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, 2014. 3, 4
- [33] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, arXiv preprint arXiv:1605.03259. 3

- [34] Z. Shi, T. M. Hospedales, T. Xiang, Transferring a semantic representation for person re-identification and search, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015. 3
- [35] S. Park, B. X. Nie, S.-C. Zhu, Attribute and-or grammar for joint parsing of human attributes, part and pose, arXiv preprint arXiv:1605.02112. 3
- [36] N. Sarafianos, B. Boteanu, B. Ionescu, I. A. Kakadiaris, 3D human pose estimation: A review of the literature and analysis of covariates, *Computer Vision and Image Understanding* 152 (2016) 1–20. 3
- [37] M. Livne, L. Sigal, N. F. Troje, D. J. Fleet, Human attributes from 3d pose tracking, *Computer Vision and Image Understanding* 116 (5) (2012) 648–660. 3
- [38] X. Xu, I. A. Kakadiaris, Joint head pose estimation and face alignment framework using global and local cnn features, in: Proc. 12th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC, Vol. 2, 2017. 3
- [39] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015. 3
- [40] J. Huang, R. S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proc. IEEE International Conference on Computer Vision, Boston, MA, 2015. 3
- [41] J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, X. Ruan, Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval, in: Proc. British Machine Vision Conference, York, UK, 2016. 3
- [42] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (3) (2014) 453–465. 3
- [43] D. Jayaraman, K. Grauman, Zero-shot recognition with unreliable attributes, in: Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014. 3
- [44] Z. Shi, Y. Yang, T. Hospedales, T. Xiang, Weakly-supervised image annotation and segmentation with objects and attributes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 3
- [45] L. Liu, A. Wiliem, S. Chen, K. Zhao, B. C. Lovell, Determining the best attributes for surveillance video keywords generation, in: Proc. IEEE International Conference on Identity, Security and Behavior Analysis, Sendai, Japan, 2016. 3
- [46] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012. 3
- [47] Y. Yang, T. M. Hospedales, A unified perspective on multi-domain and multi-task learning, arXiv preprint arXiv:1412.7489. 3
- [48] E. M. Hand, R. Chellappa, Attributes for improved attributes: A multi-task network for attribute classification, arXiv preprint arXiv:1604.07360. 3
- [49] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098. 3
- [50] A. Pentina, V. Sharmanska, C. H. Lampert, Curriculum learning of multiple tasks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015. 3
- [51] J. Yang, R. Yan, A. G. Hauptmann, Cross-domain video concept detection using adaptive svms, in: Proc. 15th ACM International Conference on Multimedia, 2007, pp. 188–197. 3
- [52] N. Sarafianos, M. Vrigkas, I. A. Kakadiaris, Adaptive SVM+: Learning with privileged information for domain adaptation, arXiv preprint arXiv:1708.09083. 3
- [53] V. Cirik, E. Hovy, L.-P. Morency, Visualizing and understanding curriculum learning for long short-term memory networks, arXiv preprint arXiv:1611.06204. 3
- [54] W. Zaremba, I. Sutskever, Learning to execute, arXiv preprint arXiv:1410.4615. 3
- [55] Q. Dong, S. Gong, X. Zhu, Multi-task curriculum transfer deep learning of clothing attributes, arXiv preprint arXiv:1610.03670. 3
- [56] M. Long, J. Wang, M. I. Jordan, Deep transfer learning with joint adaptation networks, arXiv preprint arXiv:1605.06636. 3
- [57] Y. Bengio, et al., Deep learning of representations for unsupervised and transfer learning, *ICML Unsupervised and Transfer Learning* 27 (2012) 17–36. 3
- [58] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531. 4
- [59] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector CNNs, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 4, 7, 14
- [60] D. Lopez-Paz, B. Schölkopf, L. Bottou, V. Vapnik, Unifying distillation and privileged information, in: Proc. International Conference on Learning Representations, San Jose, Puerto Rico, 2016. 4, 14
- [61] D. Jayaraman, F. Sha, K. Grauman, Decorrelating semantic visual attributes by resisting the urge to share, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014. 4
- [62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. International Conference on Learning Representations, San Diego, CA, 2015. 4
- [63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, *International Journal of Computer Vision* 115 (3) (2015) 211–252. 4
- [64] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167. 4
- [65] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research* 15 (1) (2014) 1929–1958. 4
- [66] I. Sutskever, J. Martens, G. E. Dahl, G. E. Hinton, On the importance of initialization and momentum in deep learning, in: Proc. International Conference on Machine Learning, Atlanta, GA, 2013. 4
- [67] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. Advances in neural information processing systems, Lake Tahoe, 2012. 5
- [68] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 5, 14
- [69] D. Martinho-Corbishley, M. S. Nixon, J. N. Carter, Soft biometric retrieval to describe and identify surveillance images, in: Proc. IEEE International Conference on Identity, Security and Behavior Analysis, Miyagi, Japan, 2016. 6, 7, 8
- [70] D. Adjeroh, D. Cao, M. Piccirilli, A. Ross, Predictability and correlation in human metrology, in: Proc. IEEE International Workshop on Information Forensics and Security, Seattle, WA, 2010. 5
- [71] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. 5

- [72] W. J. H. Jr., Hierarchical grouping to optimize an objective function, *Journal of the American statistical association* 58 (301) (1963) 236–244. [6](#)
- [73] J. Zhu, S. Liao, Z. Lei, S. Z. Li, Multi-label convolutional neural network based pedestrian attribute classification, *Image and Vision Computing*. [7](#), [8](#), [11](#), [13](#), [18](#)
- [74] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Rio, Brazil, 2007. [7](#), [8](#), [11](#)
- [75] Y. Deng, P. Luo, C. C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: *Proc. ACM International Conference on Multimedia*, Orlando, FL, 2014. [7](#), [8](#)
- [76] D. Parikh, K. Grauman, Relative attributes, in: *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011. [8](#)
- [77] R. Layne, T. M. Hospedales, S. Gong, Attributes-based re-identification, Springer, 2014, pp. 93–117. [8](#)
- [78] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: *Proc. International Conference on Machine Learning*, Atlanta, GA, 2013. [9](#)
- [79] S.-J. Huang, Z.-H. Zhou, Z. Zhou, Multi-label learning by exploiting label correlations locally., in: *Proc. AAAI*, Toronto, Canada, 2012. [13](#)
- [80] B. Hariharan, L. Zelnik-Manor, M. Varma, S. Vishwanathan, Large scale max-margin multi-label classification with priors, in: *Proc. International Conference on Machine Learning*, Haifa, Israel, 2010. [13](#)
- [81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015. [14](#)
- [82] R. K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, arXiv preprint arXiv:1505.00387. [14](#)

Table 6: **Appendix A - Complete Results on the PETA dataset:** Performance comparison on the PETA dataset for different types of attributes. The imbalance ratio is defined as the ratio of the number of instances in the majority class to the number of examples in the minority class in the training set. An asterisk next to an attribute denotes that it belongs to the strongly correlated group of tasks.

Visual Attribute	Imbalance Ratio	Accuracy (%)		Recall rate (%) @FPR = 10%		AUC (%)	
		Zhu <i>et al.</i> [73]	CILICIA	Zhu <i>et al.</i> [73]	CILICIA	Zhu <i>et al.</i> [73]	CILICIA
accessoryHat*	8.78	96.05	96.56	86.06	85.23	92.62	92.28
accessoryMuffler*	11.06	97.17	97.04	88.42	91.23	94.47	95.62
accessoryNothing	3.03	86.11	86.45	52.57	51.93	86.09	86.75
carryingBackpack	4.01	84.30	84.82	58.40	64.20	85.19	89.62
carryingMessengerBag	2.37	79.58	80.59	58.30	60.13	82.01	83.35
carryingNothing	2.68	80.14	81.12	55.15	61.11	83.08	85.30
carryingOther	4.11	80.91	81.69	46.90	50.68	77.68	78.16
carryingPlasticBags*	11.87	93.45	94.72	67.30	76.11	86.01	90.77
footwearBlack	1.26	75.97	75.33	57.24	55.59	84.07	83.42
footwearBrown*	13.64	92.14	92.65	65.77	62.73	85.26	85.28
footwearGrey	5.40	87.07	88.07	50.80	60.54	80.92	84.54
footwearLeatherShoes	2.39	85.26	86.16	72.28	76.79	89.84	90.96
footwearShoes	1.75	75.78	76.95	52.80	55.52	81.63	81.73
footwearSneakers*	3.67	81.78	81.08	52.04	50.63	83.19	83.43
footwearWhite*	4.73	85.89	85.73	62.72	60.44	86.16	85.26
hairBlack	1.55	87.83	87.97	81.03	82.46	93.61	94.13
hairBrown	3.86	89.58	89.07	77.36	81.22	91.33	92.48
hairGrey*	11.36	95.25	95.09	74.91	80.65	89.42	91.65
hairLong	3.23	88.12	88.21	76.49	78.93	90.55	92.15
hairShort*	2.84	86.93	87.03	69.68	69.28	89.84	90.38
lowerBodyBlack	1.08	83.86	85.41	71.21	79.42	90.84	92.83
lowerBodyBlue*	4.51	88.64	89.53	77.26	80.16	90.81	92.06
lowerBodyCasual	6.33	90.54	89.58	56.23	58.76	87.49	88.19
lowerBodyFormal	6.45	90.86	90.87	72.52	72.94	87.79	88.94
lowerBodyGrey*	3.10	82.07	81.85	53.43	51.73	82.77	83.33
lowerBodyJeans*	2.28	83.13	82.57	67.59	64.85	87.71	86.46
lowerBodyTrousers*	1.06	76.26	75.90	56.19	54.77	84.16	83.34
personalLarger60*	16.21	97.58	97.24	90.71	90.33	94.94	95.96
personalLess30	1.00	81.05	81.84	63.75	67.34	88.50	89.42
personalLess45	2.03	79.87	79.39	59.42	58.64	84.62	85.25
personalLess60*	8.96	92.84	93.21	70.22	71.94	87.66	88.03
personalMale*	1.22	84.34	86.03	74.80	77.87	91.74	92.85
upperBodyBlack	1.25	86.21	86.86	80.11	82.07	93.06	94.07
upperBodyBlue*	12.05	94.53	95.74	76.19	84.64	90.92	94.51
upperBodyBrown*	12.59	93.25	94.20	68.60	79.80	87.58	92.54
upperBodyCasual*	5.98	89.25	80.00	62.14	61.45	87.17	85.08
upperBodyFormal	6.66	91.12	90.33	70.48	74.47	87.57	89.60
upperBodyGrey	4.56	84.39	84.46	55.33	63.38	82.99	87.44
upperBodyJacket*	13.20	92.34	93.04	53.37	63.10	80.98	85.27
upperBodyLongSleeve	4.98	87.88	87.96	74.29	80.69	89.97	92.72
upperBodyOther*	1.22	81.97	81.58	73.19	71.83	88.50	88.14
upperBodyRed*	17.66	96.33	96.64	86.77	90.82	94.69	96.16
upperBodyShortSleeve	5.98	88.09	88.79	69.22	75.88	89.21	92.10
upperBodyTshirt*	10.53	90.59	90.80	63.51	68.65	88.73	90.55
upperBodyWhite	18.87	88.84	90.37	75.25	84.18	91.24	94.65
Strongly Cor. Av.	8.12	89.63	89.87	70.04	72.19	88.42	89.50
Weakly Cor. Av.	4.12	84.93	85.32	64.66	68.56	86.93	88.60
Total Av.	6.07	87.23	87.54	67.29	70.34	87.66	89.04