

# Curriculum Learning for Multi-Task Classification of Visual Attributes

Nikolaos Sarafianos<sup>1</sup>  
 Christophoros Nikou<sup>3</sup>

Theodore Giannakopoulos<sup>2</sup>  
 Ioannis A. Kakadiaris<sup>1</sup>

<sup>1</sup>University of Houston    <sup>2</sup>NCSR Demokritos    <sup>3</sup>University of Ioannina

## 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 based on their correlation so that two groups of strongly and weakly correlated tasks are formed. The two groups of tasks are learned in a curriculum learning setup by transferring the acquired knowledge from the strongly to the weakly correlated. The learning process within each group though, is performed in a multi-task classification setup. The proposed method learns better and converges faster than learning all the tasks in a typical multi-task learning paradigm. We demonstrate the effectiveness of our approach on the publicly available, SoBiR, VIPeR and PETA datasets and report state-of-the-art results across the board.

## 1. Introduction

Moments after the Boston marathon bombing, the FBI gathered almost 10TB of photos and videos, looking for a “backpack-carrying man, wearing a white hat”. In suspect descriptions, humans tend to rely on visual attributes since (i) they can be composed in different ways to create descriptions; (ii) they are generalizable as with some fine-tuning they can be applied to recognize objects for different tasks; and (iii) they are a meaningful semantic representation of objects or humans that can be understood by both computers and humans. Given an image of a human, a question that arises is how can someone effectively predict the corresponding visual attributes?

In this work, we propose CILICIA (Curriculum Learning multiTask Classification Attributes) to address the problem of visual attribute classification from images of humans. Instead of using low-level representations which

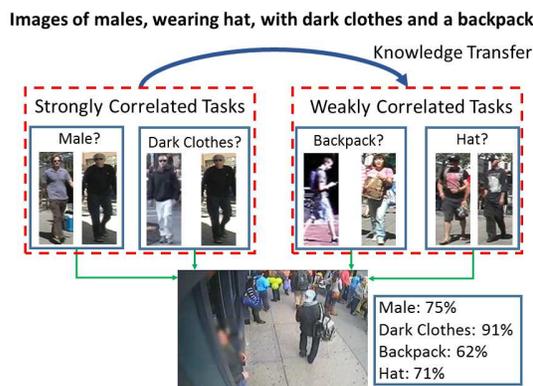


Figure 1: Can we do better in visual attribute multi-task classification? Wouldn't it be great if we could find a way to learn the attributes in a more semantically meaningful way instead of all at the same time? Our approach aspires to combine the advantages of curriculum learning and multi-task classification to predict the visual attributes of humans.

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 two groups based on their cross-correlation. The group of the strongly correlated attributes is learned first, and then the acquired knowledge is transferred to the second group.

When Vapnik and Vashist introduced the learning using privileged information (LUPI) paradigm [31], 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 [15, 20, 25, 27, 32, 33].

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 way as in the method of Yan *et al.* [34] or with a curriculum strategy. Curriculum learning [2, 14] (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 into strongly and weakly correlated. In that way, we exploit the advantages of both multi-task and curriculum learning. First, the ConvNet learns the strongly correlated tasks 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 two groups based on their correlation with the rest of the tasks. The tasks of each subgroup are learned in a joint manner. Thus, the proposed method learns better and converges faster than learning all the tasks in a typical multi-task learning setup. Second, we propose a scheme of transferring knowledge between the groups of tasks which reduces the convergence time and increases the performance. We performed extensive evaluations, ablation studies and an analysis of the covariates in one small-scale dataset and one medium-scale dataset and achieved state-of-the-art results.

## 2. Related Work

**Visual Attributes:** Predicting the visual attributes of a human from an image is not a new concept as it has previously been addressed in the literature in many contexts. Ferrari and Zisserman [6] were the first to investigate the power of visual attributes. They used low-level features and a probabilistic generative model to learn these attributes and segment them in an image. Kumar *et al.* [17] proposed an automatic method to perform face verification and image search by training classifiers for describable facial visual attributes (e.g., gender, hair color, and eyewear). Scheirer *et al.* [26]

proposed a novel method to construct normalized “multi-attribute spaces” from raw classifier outputs. However, they focused entirely on the score calibration without investigating the feature extraction part. Following the deep learning renaissance, several papers [7, 8, 18] have addressed the visual attribute classification problem using ConvNets. Zhang *et al.* [37] proposed an attribute classification method which combines part-based models in the form of poselets [3], and deep learning by training pose-normalized ConvNets. Their method though, requires training a network for each poselet which is a computationally expensive task. Zhu *et al.* [39] 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, learning them jointly will be beneficial. Finally, a very interesting prior work which focuses on the correlation of visual attributes is the method of Jayaraman *et al.* [13]. While our work also leverages information from correlated attributes in a multi-task classification framework, it models co-occurrence between different groups of visual attributes instead of trying to semantically decorrelate them.

**Curriculum Learning:** Solving all tasks jointly is commonly employed in the literature [4, 10, 39] 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 [23]. However, some tasks are easier than others and also not all tasks are equally related to each other [22]. Curriculum Learning was initially proposed by Bengio *et al.* [2]. 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.* [22] 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. 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.

## 3. Methodology

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 parts of the network that solve the strongly and the weakly correlated tasks

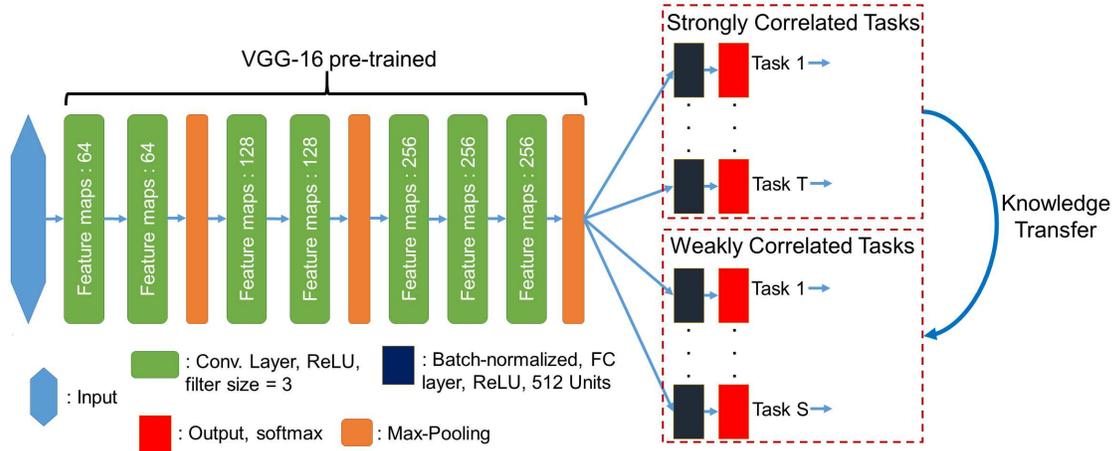


Figure 2: Architecture of the ConvNet used in our framework for both strongly and weakly correlated tasks. The VGG-16 pre-trained part is kept frozen during training and only the weights of the last layers are learned. The two parts are learned separately. However, when the weakly correlated tasks are trained, both tasks contribute to the total cost function.

as  $C_s$  and  $C_w$ , respectively.

### 3.1. Multi-label ConvNet

To mitigate the lack of training data we employ the pre-trained VGG-16 [28] network. VGG-16, is the network 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. VGG-16 is trained on ImageNet [24], 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 [35] and thus, not transferable. Following that, for every task we added a batch-normalized [12] 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 [29] 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 backpropagated 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.* [30], 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 0.0001 for the layers that are trained; (iii) initialized the learning rate at 0.001 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 pixel values as proposed by Krizhevsky *et al.* [16]. 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.* [11], proved to be very effective as it reduced the variation on the predictions.

### 3.2. Correlation-based Group Split

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:

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

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 task  $j$ , and  $\mathcal{E}$  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.* [1] 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 others. The correlation between different sub-problems was also exploited in the age estimation method of Niu *et al.* [21] in an ordinal regression setup.

To address this problem we propose to find the total dependency  $p_i$  of task  $t_i$  with the rest, by computing the respective Pearson correlation coefficients:

$$p_i = \sum_{j=1, j \neq i}^T \frac{\text{cov}(y_{t_i}, y_{t_j})}{\sigma(y_{t_i})\sigma(y_{t_j})}, \quad i = 1, \dots, T \quad (2)$$

where  $\sigma(y_{t_i})$  is the standard deviation of the labels  $y$  of the task  $t_i$ . After we compute the total dependencies for all attributes, the obtained vector of size  $T \times 1$  (each value corresponds to one line of the Pearson correlation coefficient matrix) is sorted in a descending order. Tasks with a top 50% of  $p_i$  are strongly correlated with the rest, and thus they are assigned to the strongly correlated group. The remaining tasks are assigned as weakly correlated and will employ the information learned from the former group.

### 3.3. Multi-Task Curriculum Learning

In the scenario we are investigating, we solve multiple binary unbalanced classification tasks simultaneously. Thus, similar to Zhu *et al.* [38] we employ the categorical cross-entropy function between predictions and targets, which for a single attribute  $t$  is defined as follows:

$$L_t = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M \left( \frac{1/M_j}{\sum_{n=1}^M 1/M_n} \right) \cdot \mathbb{1}[y_i = j] \cdot \log(p_{i,j}), \quad (3)$$

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.

However, in the method of Zhu *et al.* [38] the total loss over all attributes is defined as  $L_s = \sum_{t=1}^T \lambda_t \cdot L_t$ , where  $\lambda_t$  is the contribution weight of each parameter. For simplicity, it is set to  $\lambda_t = 1/T$ , but this is problematic since there is

---

#### Algorithm 1: Multi-task curriculum learning training

---

**Input** : Training set  $X$ , training labels  $Y$

- 1  $Y_s, Y_w \leftarrow$  using the observations  $X$ , split labels  $Y$  by maximizing Eq. (2)
- 2  $C_s \leftarrow$  freeze  $C_w$ , train model using  $(X, Y_s)$  by minimizing the loss in Eq. (3)
- 3 Initialize  $C_w$  from  $C_s$
- 4  $C_w \leftarrow$  train model using  $(X, Y_w)$  by minimizing the loss in Eq. (4)

**Output**: Parameters of networks  $C_s$  and  $C_w$  for the strongly and the weakly correlated tasks, respectively

---

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 the difference in the performance was statistically insignificant, and thus we did not investigate this any further.

Once the classification of the visual-attribute tasks that demonstrated a strong correlation with the rest is performed, we use the learned parameters (*i.e.*, weights, biases and batch normalization parameters) to initialize the network for the less diverse attributes. Its architecture remains the same, with the parameters of VGG-16 being kept “frozen”. When the number of tasks is odd, then an additional “branch” is added at the end of the network to learn the task-specific parameters. Furthermore, by adopting the “supervision transfer” technique of Zhang *et al.* [36] we leverage the knowledge learned by backpropagating the following loss:

$$L_w = \lambda \cdot L_s + (1 - \lambda) \cdot L_w^f, \quad (4)$$

where  $L_w^f$  is the total loss computed during the forward pass using Eq. (3) only over the weakly correlated tasks and  $\lambda$  is a parameter that controls the amount of knowledge transferred. Throughout our experimental investigation we found that a 25% contribution of the already learned group of strongly correlated tasks yielded the best results.

The process of computing the two groups of attributes 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 split, it is not computationally intensive. Finally, note that, the group split depends on the training set and it’s possible that different train-test splits might yield different groups of tasks which is why average classification results are reported over five random splits.

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. Attributes are in descending order based on their cross-correlation. Those in the second group correspond to the weakly correlated.

Soft Label	SVM	Individual Learning	Multi-Task Learning	CILICIA
Weight	57.7	67.7	71.0	<b>73.6</b>
Figure	57.8	68.7	68.6	<b>71.8</b>
Muscle build	58.5	73.3	<b>74.5</b>	73.6
Arm thickness	60.1	72.0	<b>73.1</b>	70.7
Leg thickness	56.7	68.9	71.0	<b>73.0</b>
Chest size	58.7	64.9	68.9	<b>70.7</b>
Age	58.5	<b>62.6</b>	61.9	59.7
Height	64.7	73.9	72.0	<b>75.7</b>
Skin color	59.2	66.8	<b>68.0</b>	67.8
Hair color	67.5	74.2	78.1	<b>78.5</b>
Hair length	71.8	78.9	79.2	<b>79.6</b>
Gender	72.1	<b>81.4</b>	79.6	81.3
Strongly Cor.	58.3	69.3	71.3	<b>72.3</b>
Weakly Cor.	65.6	73.0	73.2	<b>73.7</b>
Total Av.	61.9	71.2	72.3	<b>73.1</b>

## 4. Experiments

### 4.1. Results on SoBiR

Since the SoBiR dataset [19] does not have a baseline on attribute classification we reported results using hand-crafted 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  $128 \times 128$ . 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

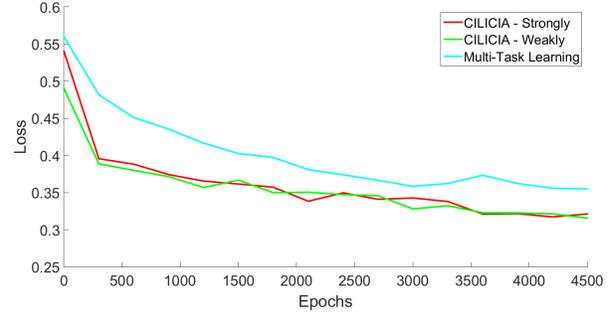


Figure 3: Convergence plot for both groups of CILICIA and Multi-Task learning on the SoBiR dataset. Note that the first group corresponds to the strongly correlated and the second to the weakly correlated group of tasks.

Table 2: Performance comparison on the VIPeR dataset. Attributes are in descending order based on their cross-correlation. Those in the second group correspond to the weakly correlated.

Visual Attribute	Multi-Task Learning	Zhu <i>et al.</i> [39]	CILICIA
barelegs	79.6 ± 0.8	<b>84.1</b> ± 1.1	82.9 ± 0.7
shorts	76.8 ± 1.1	81.7 ± 1.3	<b>85.2</b> ± 0.3
nocoats	74.3 ± 1.3	71.3 ± 0.8	<b>71.3</b> ± 0.5
skirt	67.2 ± 3.7	78.1 ± 3.5	<b>86.2</b> ± 3.8
nolightdarkjeanscolor	87.1 ± 1.6	90.7 ± 2.0	<b>96.7</b> ± 0.4
redshirt	79.2 ± 1.9	91.9 ± 1.0	<b>95.1</b> ± 0.4
patterned	67.4 ± 3.5	57.9 ± 9.2	<b>77.5</b> ± 4.3
hashandbag	66.9 ± 3.1	42.0 ± 6.5	<b>81.5</b> ± 2.7
greenshirt	70.3 ± 2.4	75.9 ± 5.9	<b>90.5</b> ± 2.3
lightshirt	79.5 ± 0.9	83.0 ± 1.2	<b>84.0</b> ± 0.8
blueshirt	69.9 ± 1.7	69.1 ± 3.3	<b>90.2</b> ± 0.7
lightbottoms	<b>79.0</b> ± 1.0	<b>76.4</b> ± 1.2	72.5 ± 0.4
hassatchel	72.5 ± 0.8	57.8 ± 2.7	<b>72.8</b> ± 0.3
midhair	74.3 ± 1.3	76.1 ± 1.8	<b>77.6</b> ± 1.4
male	71.5 ± 1.9	69.6 ± 2.6	<b>71.5</b> ± 1.2
darkhair	70.1 ± 2.0	<b>73.1</b> ± 2.1	64.9 ± 1.2
hasbackpack	68.4 ± 1.4	64.9 ± 1.2	<b>70.2</b> ± 0.4
darkbottoms	68.1 ± 0.9	<b>78.4</b> ± 0.7	75.2 ± 0.8
jeans	74.9 ± 0.7	<b>77.5</b> ± 0.6	74.9 ± 0.6
darkshirt	71.0 ± 1.4	82.3 ± 1.4	<b>84.3</b> ± 0.5
Strongly Cor. Av.	73.4 ± 2.6	75.7 ± 3.2	<b>85.1</b> ± 1.0
Weakly Cor. Av.	71.9 ± 1.8	72.5 ± 1.7	<b>74.8</b> ± 0.5
Total Av.	73.2 ± 1.2	74.1 ± 1.0	<b>80.5</b> ± 0.7

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 62.6% when it is learned individually using our ConvNet architecture. Finally for completeness and to demonstrate the convergence of all learning schemes, we provide in Figure 3 the convergence plots for both CILICIA groups and Multi-Task learning.

## 4.2. Results on VIPeR

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.* [39] and a typical multi-task learning framework using the VIPeR dataset [9]. 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 6.4%. Our method is superior in both groups but especially in the strongly correlated group of labels, in which the improvement is almost 10%. CILICIA achieved 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 one to nine) due to the way annotation is provided (*e.g.*, when the question is “is the human wearing a red t-shirt or not” the answer is mainly negative).

## 5. Performance Analysis and Ablation Studies

The proposed approach outperformed the state-of-the-art in all three datasets. We argue that the main reasons for this are: (i) we exploited the correlation between different attributes and learned a model to classify them in two steps; (ii) the knowledge transfer from the strongly correlated to the weakly correlated attributes which improved the performance and reduced the required training time; and (iii) the use of a pre-trained deep architecture with the first layers frozen which was not the case in the method of Zhu *et al.* [39]. To assess the impact of both contributions and to demonstrate their effectiveness we conducted two ablation studies. 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.

**Effectiveness of knowledge transfer:** In the first ablation study we compare the classification accuracy of the selected tasks with and without knowledge transfer. When no knowledge is transferred we are simply training two multi-task classification frameworks. We report the obtained results in the last two columns of Table 3. Transferring knowledge from the strongly to the weakly correlated group of tasks improves the performance of the latter by 1.89% compared to a typical multi-task classification learning framework.

**Effectiveness of correlation-based split:** In the second study, we use the same eight selected attributes but instead of grouping them based on their cross-correlation, we

Table 3: Ablation experiments to assess the effectiveness of knowledge transfer and correlation-based split using the four most and the four least correlated attributes of the PETA dataset. 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) refers 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	<b>71.80</b>
Total	64.22	72.95	<b>73.91</b>

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 3. 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 completely different) has an adverse effect on the performance.

## 6. Conclusion

In this paper, we introduced CILICIA, a multi-task curriculum learning method to address the visual-attribute classification problem. Given images of humans as an input, we performed end-to-end learning by solving multiple binary classification problems simultaneously. Tasks were grouped based on their cross-correlation so that two groups of strongly and weakly correlated tasks are formed. The attributes of each group are then learned in a multi-task learning setup. During training of the weakly correlated tasks, we leveraged the knowledge already learned from the strongly correlated tasks. 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. 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] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross. Predictability and correlation in human metrology. In *Proc. IEEE International Workshop on Information Forensics and Security*, Seattle, WA, Dec. 12-15 2010. 4
- [2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. International Conference on Machine Learning*, Montreal, Canada, June 14-18 2009. 2
- [3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6-13 2011. 2
- [4] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16-21 2012. 2
- [5] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM International Conference on Multimedia*, Orlando, FL, Nov. 3-7 2014.
- [6] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 3-6 2007. 2
- [7] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 7-12 2015. 2
- [8] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with R\*CNN. In *Proc. IEEE International Conference on Computer Vision*, Boston, MA, June 7-12 2015. 2
- [9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Rio, Brazil, Oct. 14 2007. 6
- [10] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360*, 2016. 2
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 3
- [12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3
- [13] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, June 23-28 2014. 2
- [14] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 8-13 2014. 2
- [15] I. A. Kakadiaris, N. Sarafianos, and C. Nikou. Show me your body: Gender classification from still images. In *IEEE International Conference on Image Processing*, Phoenix, AZ, Sept. 24-28 2016. 1
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in neural information processing systems*, Lake Tahoe, Dec. 3-8 2012. 3
- [17] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011. 2
- [18] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *Proc. European Conference on Computer Vision*, Amsterdam, The Netherlands, Oct. 8-16 2016. 2
- [19] D. Martinho-Corbishley, M. S. Nixon, and 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, March 1-2 2016. 5
- [20] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Information bottleneck learning using privileged information for visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 1
- [21] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output CNN for age estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 4
- [22] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, June 7-12 2015. 2
- [23] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3
- [25] N. Sarafianos, C. Nikou, and I. Kakadiaris. Predicting privileged information for height estimation. In *Proc. International Conference on Pattern Recognition*, Cancun, Mexico, Dec. 4-8 2016. 1
- [26] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and 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, June 16-21 2012. 2
- [27] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3-6 2013. 1
- [28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations*, San Diego, CA, May 7-9 2015. 3
- [29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3
- [30] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton. On the importance of initialization and momentum in deep learning.

- In *Proc. International Conference on Machine Learning*, Atlanta, GA, June 16-21 2013. 3
- [31] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–57, 2009. 1
- [32] M. Vrigkas, C. Nikou, and I. A. Kakadiaris. Active privileged learning of human activities from weakly labeled samples. In *IEEE International Conference on Image Processing*, Phoenix, AZ, Sept. 24-28. 1
- [33] S. Wang, D. Tao, and J. Yang. Relative attribute SVM+ learning for age estimation. *IEEE Transactions on Cybernetics*, 46(3):827–839, 2015. 1
- [34] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. HD-CNN: Hierarchical deep convolutional neural network for large scale visual recognition. In *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, Dec. 13-16 2015. 2
- [35] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 8-13 2014. 3
- [36] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector CNNs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016. 4
- [37] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, June 23-28 2014. 2
- [38] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 2016. 4
- [39] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *Proc. IEEE International Conference on Biometrics*, Phuket, Thailand, May 19-22 2015. 2, 5, 6