PERSON SEARCH USING IDENTITY ATTRIBUTES

A Dissertation Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Nikolaos Sarafianos

May 2019

PERSON SEARCH USING IDENTITY ATTRIBUTES

Nikolaos Sarafianos

APPROVED:

Ioannis A. Kakadiaris, Chairman
Dept. of Computer Science

Ioannis Pavlidis
Dept. of Computer Science

Guoning Chen
Dept. of Computer Science

Omprakash Gnawali
Dept. of Computer Science

Theodoros Giannakopoulos
NCSR Demokritos

Dean, College of Natural Sciences and Mathematics

ACKNOWLEDGMENTS

Through the course of this 5-year journey, there are a lot of people who helped me, believed in me and contributed to improving me personally and professionally. First and foremost I would like to thank my academic advisor, Prof. Ioannis A. Kakadiaris, who believed in me when he accepted me at CBL. He gave me endless opportunities to grow, taught me how to do research and how to work on the highest standards. He invested tons of time and resources on me to ensure that I have everything I need to grow and learn and he has helped me immensely to become a better researcher as well as a better person. I would like to thank my Ph.D. committee members namely Prof. Ioannis Pavlidis, Prof. Guoning Chen, Prof. Omprakash Gnawali, and Dr. Theodoros Giannakopoulos for their great support and invaluable advice. Their domain expertise, the great quality of their classes that I attended as well as their constructive feedback during the course of my studies have been invaluable. I would also like to thank colleagues that I worked with and learned from: Dr. Theodoros Giannakopoulos, Prof. Christophoros Nikou, Prof. Bogdan Ionescu, Dr. Michalis Vrigkas, Dr. Bogdan Boteanu, and Xiang Xu. By working with them and publishing papers together I got to learn from them and I am very thankful for their guidance. I would also like to thank my managers and mentors during my internships who believed in me and gave me the opportunity to work in the industry during my studies. I would like to thank Dr. Chao Wang and Dr. Shiva Sundaram from Amazon as well as Dr. Tony Tung from Facebook Reality Labs. I would like to thank all of my friends at CBL past and present that helped make this journey really enjoyable. I would like to thank Rachel Reed for all her support these past few years. Finally, I would like to thank my

family who has believed in me from day one and has been there every single time I needed them. This dissertation is in the memory of Jonathan Kovar.

PERSON SEARCH USING IDENTITY ATTRIBUTES

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Nikolaos Sarafianos

May 2019

Abstract

The goal of this dissertation is to develop and evaluate algorithms and a prototype system to retrieve frames depicting humans with specific identity attributes obtained from a textual description. Solving this problem requires addressing three separate subproblems, namely (i) defining the ontology of the identity and the identity-related attributes, (ii) developing and evaluating algorithms for extracting identity attributes from images, and (iii) developing and evaluating an algorithm for attribute-based person search in databases of image frames. This dissertation presents a list of methods on visual attribute classification and person search that significantly improve the accuracy over previous work. The methods presented tackle key limitations of previous work such as the class imbalance of visual attributes, or the challenge of learning discriminative representations from the textual input. By learning to retrieve the most relevant images of individuals based on textual descriptions, such techniques can have a broader impact in cases of missing children or in surveillance applications. The works introduced in this dissertation are capable of successfully identifying which images contain humans with such characteristics which could reduce dramatically the effort and the time required to identify such information. In each method a detailed overview of the benefits and limitations of each approach is introduced, extensive experimental evaluation and ablation studies are provided to analyze the impact of different modules, and further limitations have been identified that need to be addressed by future work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

When humans are asked to provide a description of an object or a human, they tend to use visual attributes to accomplish this task. For example, a laptop can have a widescreen, 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. Humans 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 in search for a product online; (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 because images might be of low-resolution, humans might be partially occluded in cluttered scenes, or there might be significant pose variations.

1.2 Unsolved Challenges

Despite the promising performance of several person search and visual attribute classification algorithms in the literature, there are several challenges that still remain.

Class Imbalance: Human attributes are imbalanced in nature. Bald people with a mustache wearing glasses are 14 to 43 times less likely to appear in the CelebA dataset [73] compared to people without these characteristics. Large-scale imbalanced datasets can lead to biased models, optimized to favor the majority classes while failing to identify the subtle discriminant features that are required to recognize the under-represented classes.

Attribute Presence in Images: An additional challenge is identifying which areas in the image provide class-discriminant information. Giving emphasis to the upper part of an image, where the face is located, for attributes such as "glasses" and to the bottom part for attributes such as "long pants" can increase the recognition performance as well as the interpretability of the designed models [85]. This challenge is usually addressed using visual attention techniques that output saliency maps. However, in the human attribute domain, attention ground-truth annotations are not available to learn such spatial attributions.

Learning from Text: Textual descriptions contain a large variability of words even when they are used to describe the same image. What is considered as important information for

one is not necessarily the same for another annotator. At the same time, textual descriptions might contain obvious mistakes, the descriptions can be too long or the annotator might describe additional information that is available on the image but is not related to the person of interest. All these factors make the text-to-image retrieval a difficult problem since learning good feature representations from such descriptions is not straightforward.

1.3 Limitations of Previous Work

Class Imbalance: Learning from imbalanced data is a well-studied problem in machine learning. Traditional solutions include over-sampling the minority classes [10, 77] or under-sampling the majority classes [23] to compensate for the imbalanced class ratio and cost-sensitive learning [52] where classification errors are penalized differently. Such approaches have been extensively used in the past but they suffer from some limitations. For example, over-sampling introduces redundant information making the models more over-fitting-prone, whereas under-sampling may remove valuable discriminative information. Recent works with deep convolutional neural networks [39, 21] introduced a sampling procedure of quintuplets or triplets of samples that satisfy some properties in the feature-space and used them to regularize their models. However, sampling triplets is a computationally expensive procedure and the characteristics of the triplets in a batch-mode setup might vary significantly.

Attribute Presence in Images: Modern visual attribute classification techniques rely either on contextual information [67, 29], side information [99], curriculum learning strategies [96] or visual attention mechanisms [142] to accomplish their task. Although context

and side information can increase the recognition accuracy, such approaches are over-complicated and difficult to train or require additional annotations that might not be available.

Learning from Text: Setting the visual attributes aside, person search using textual descriptions which is the third objective of this dissertation has received significant attention recently due to the CUHK-PEDES dataset that was published in 2017. However must approaches suffer from important limitations: (i) they are constrained by the existing learning objective functions, (ii) they seem to ignore the large variability of the textual input by solely relying on an LSTM to model the input sentences, and (iii) they demonstrate subpar results as the best text-to-image rank-1 accuracy results in the literature are below 50% and 40% in the CUHK-PEDES and Flickr30K datasets, respectively.

1.4 Goal and Objectives

The goal of this dissertation is to develop and evaluate algorithms and a prototype system to search in a database of frames depicting humans with specific identity attributes provided by the textual descriptions. Identity attributes are the set of traits used to describe a human that can be further split to identity attributes such as gender, height, age as well as identity-related attributes such as clothing and accessories. Solving this problem requires addressing two separate subproblems, namely visual attribute classification and attribute-based person search. In this thesis, new methods that address the challenges and limitations are proposed. In particular, the objectives of this dissertation are to:

1. Define an ontology of the identity and identity-related attributes. This refers to finding the set of soft-biometrics, visual attributes, scene attributes and other meta-data that can describe an image of a human and can be used to perform person search in videos.

2. Develop and evaluate an algorithm for extracting identity attributes from images. Such algorithms can be further separated into three subcategories: namely 3D human-pose estimation, soft-biometric prediction using privileged information and deep-visual attribute classification.

3. Develop and evaluate an algorithm for attribute-based search in databases of images.

1.5 Contributions

1.5.1 Major Contributions

The major contributions of this dissertation are the following:

1. An attribute ontology is designed, implemented and evaluated which comprises identity and identity-related attributes and can help the network learn better feature representations. The impact of the ontology is evaluated on text-based person search applications and performance improvements of 2% over the previous work are obtained. (Objective 1)

2. An attribute-prediction algorithm is designed, implemented, and evaluated that handles class imbalance and utilizes visual-attention mechanisms to predict visual attributes. By addressing the challenges of class imbalance and the lack of semantic annotations, state-of-the-art results were obtained in the two most widely-used publicly available datasets. (Objective 2)

3. A text-to-image matching method is designed, implemented and evaluated that retrieves the most relevant images given a textual description as an input. By leveraging adversarial-domain training and deep-language models, better representations can be learned. Improvements ranging from 2% to 5% are obtained in a variety of different images containing humans, scenes, birds, and flowers. (Objective 3)

1.5.2 Additional Contributions

In addition to the major contributions, this dissertations has the following contributions:

1. A dataset of textual descriptions from the University of Houston police department is collected. The proposed ontology is evaluated against it to ensure that it covers most of the descriptions provided. (Objective 1)

2. An algorithm named CILICIA is designed, implemented, and evaluated that combines curriculum learning and multi-task learning in a deep framework and predicts visual attributes. By introducing a curriculum, the groups of visual attributes are learned based on their difficulty which results in increased classification accuracy in three publicly available datasets. (Objective 2)

3. A survey on 3D human-pose estimation is designed, conducted, and presented. A synthetic dataset is created to evaluate different algorithms and analyze the impact of different covariates in the final performance. (Objective 2)

1.6 Dissertation Outline

The rest of the dissertation is organized as follows: the background and related work are presented in Chapter 2. The proposed methods for each of the objectives are discussed and evaluated in Chapter 3 to Chapter 5, respectively. Finally, Chapter 6 concludes all the works and provides directions for future research.

1.7 Publications

1. N. Sarafianos, and I.A. Kakadiaris. "Text-to-Image Person Search: From Textual Descriptions to Visual Attributes," IEEE Transactions on Biometrics, Behavior, and Identity Science (under review)

2. N. Sarafianos, X. Xu and I.A. Kakadiaris. "Adversarial Representation Learning for Text-Image Cross-Domain Matching," In Proc. International Conference on Computer Vision, Seoul South Korea, Oct 8-14, 2019. (under review)

3. X. Xu, N. Sarafianos, and I.A. Kakadiaris. "On Improving the Generalization of Face Recognition in the Presence of Occlusions," In Proc. International Conference on Computer Vision, Seoul South Korea, Oct 8-14, 2019. (under review)

4. N. Sarafianos, X. Xu and I.A. Kakadiaris. “Deep Imbalanced Attribute Classification using Visual Attention Aggregation,” In Proc. European Conference on Computer Vision, Munich Germany, Sep 8-14, 2018.

5. N. Sarafianos, T. Giannakopoulos, C. Nikou, and I.A. Kakadiaris. “Curriculum Learning for Multi-Task Classification of Visual Attributes,” In Proc. International Conference on Computer Vision Workshops, Venice, Italy, Oct. 22-29, 2017

6. N. Sarafianos, M. Vrigkas, and I.A. Kakadiaris. “Adaptive SVM+: Learning with Privileged Information for Domain Adaptation,” In Proc. International Conference on Computer Vision Workshops, Venice, Italy, Oct. 22-29, 2017

7. N. Sarafianos, C. Nikou, and I.A. Kakadiaris. “Predicting Privileged Information for Height Estimation,” In Proc. International Conference on Pattern Recognition, Cancun, Mexico, Dec. 4-8, 2016

8. I.A. Kakadiaris, N. Sarafianos, and C. Nikou. “Show me your body: Gender classification from still images,” In Proc. International Conference on Image Processing, Phoenix AZ, Sep. 25-28, 2016

9. N. Sarafianos, T. Giannakopoulos, C. Nikou, and I.A. Kakadiaris. “Curriculum Learning of Visual Attribute Clusters for Multi-Task Classification,” Pattern Recognition, 80: 94-108, 2018

10. N. Sarafianos, B. Boteanu, B. Ionescu and I.A. Kakadiaris. “3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates,” Computer Vision and Image Understanding, 152: 1-20, 2016

Chapter 2

Background and Related Work

In this chapter, an overview of essential concepts and existing literature in visual attribute classification, person search, and other related fields is offered.

2.1 Visual Attributes

The first to investigate the power of visual attributes were Ferrari and Zisserman [25]. 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.* [57] 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.* [100], 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 [24, 124, 139] which select attributes based on specific criteria (e.g., entropy). Zheng *et al.*[139] formulated attribute selection as a submodular optimization problem and defined a novel submodular objective function.

Following the deep learning renaissance in 2012, several papers [67, 30, 99, 142, 21] 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.* [137] proposed an attribute classification method which combines part-based models in the form of poselets [6] and deep learning by training pose-normalized ConvNets. Gkioxari *et al.* [29] proposed a deep version of poselets to detect human body parts which were then employed to perform action and attribute classification. Zhu *et al.* [144] 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. 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 [101, 133].

Finally, visual attributes have been employed recently for re-identification [109, 110], pose estimation [95], 3D pose tracking [74], attribute mining or retrieval for clothing applications [41, 106], zero-shot visual object categorization or recognition [58], and image annotation and segmentation [103].

2.2 Visual Attention

Visual attention can be interpreted as a mechanism of guiding the network to focus its resources on those spatial parts that contain information relevant to the input image. In computer-vision applications, visual attribution is usually implemented as a gating function represented with a sigmoid activation or a spatial softmax and is placed on top of one or more convolutional layers with small kernels extracting high-level information. Several interesting works have appeared recently that demonstrate the efficiency of visual attention [142, 121, 66, 119, 16, 12]. For example, the harmonious attention of Li *et al.* [66] consists of four subparts that extract hard-regional attention, soft-spatial, and channel attention to perform person re-identification. Deciding where to place the attention mechanism in the network is a topic of active research with several single-scale and multi-scale attention techniques in the literature. Das *et al.* [19], opted for a single attention module, whereas others [121, 16] extract saliency heatmaps at multiple-scales to build richer feature representations.

2.3 Deep Imbalanced Classification

Two works that address this problem in an attribute classification framework are the large margin local embedding (LMLE) method [39] and the class rectification loss (CRL) [21]. In LMLE, quintuplets were sampled that preserve locality across clusters and discrimination between classes and a loss was introduced. Dong *et al.* [21] demonstrated that careful hard mining of triplets within the batch, acts as an effective regularization that improves the recognition performance of imbalanced attributes. However, LMLE is prohibitively computationally expensive as it comprises an alternating scheme for cluster refinement and classification. CRL on the other hand, samples triplets within the batch which complicates significantly the training process as the convergence and the performance heavily rely on the triplet selection. In addition, CRL adds a fully-connected layer for each attribute before the final classification layer, which increases significantly the number of parameters that need to be learned. For example, adding a fully-connected layer with 64 units after the last convolutional layer of a ResNet-101 introduces an additional $2048 \times 7 \times 7 \times 64 = 1.8 * 10^6$ parameters per attribute. Both methods approach class imbalance purely as a machine learning problem without focusing on the visual traits of the images that correspond to these attributes. Class imbalance arises also in detection problems [121, 70], where the foreground object (or face) covers a small part of the image. A simple yet very effective solution is focal loss [70] which uses a weighting scheme at an instance-level within the batch to penalize hard misclassified samples and assign near-zero weights to easily classified samples.

2.4 Curriculum Learning

Solving all tasks jointly is commonly employed in the literature [17, 35, 81] as it is fast, easy to scale and achieves good generalization. However, some tasks are easier than others and also not

all tasks are equally related to each other [89]. Curriculum Learning was initially proposed by Bengio *et al.* [5]. 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.* [89] 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 [132, 97]. 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 success on performing data regularization on models trained on corrupted labels [50], long short-term memory (LSTM) networks [38], reinforcement learning [80, 26], robot learning policies [82] as well as object detection [65]. In parallel with our work, Dong *et al.* [22] also proposed a multi-task curriculum transfer technique to classify clothes based on their attributes. They approached the problem in a domain adaptation setup in which a classifier is first learned on easy clean samples (source domain) and then it is adapted to harder samples (cross-domain). However, the curriculum they utilize (which images correspond to the source domain and which to the cross-domain) is selected manually based on the dataset whereas in our proposed framework it is done automatically based on the label-cross correlation before training starts.

2.5 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 [75, 4]. Zhang *et al.* [136] 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.* [76] 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.* [48]. 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.

2.6 Learning Using Privileged Information

In 2009, the learning using privileged information (LUPI) framework was introduced by Vapnik and Vashist [116]. This new paradigm places a nontrivial teacher who provides additional information (i.e. features) during the training process, but it is not available for test examples. It can be applied to both classification (i.e. SVM+ algorithm) and regression tasks (i.e. ϵ-SVR+ algorithm). Following this work, new approaches that leverage privileged information in different ways have been introduced. In the work of Sharmanska *et al.* [102], samples are examined whether they are easy or difficult to classify in the privileged space. This information (i.e. distance from the margin) is then transferred to the observable space to improve the prediction performance. Lapin *et al.* [60] related the privileged information framework to the importance of sample weighting and

showed that prior knowledge can be encoded using weights in a regular support vector machine. Recently, the LUPI paradigm was employed with applications on biometrics [131, 122, 51] such as face verification, person identification, age estimation, and gender classification.

2.7 Soft Biometrics

Integrating soft biometrics such as gender, height, weight, age, and ethnicity to a primary biometrics system (e.g., face) has been studied by Jain *et al.* [47]. In most of the existing literature [117, 78], the problem of human classification assisted by soft biometrics has been approached using facial information. However, in real-life scenarios, such information might not be available (e.g., the face might be covered or occluded). This led to methods that employ information from the human body to perform human identification and tracking based on soft biometrics [126, 45, 14]. Adjeroh *et al.* [1] studied the correlation of several anthropometric measurements from the CAESAR anthropometric database [94] and proposed a cluster-driven prediction model which employs information from human metrology. In the work of Guo *et al.* [7], the same dataset was used and a method that predicts the gender and the weight was proposed.

2.8 Person Search

Text-based Person Search: Learning cross-modal embeddings has numerous applications ranging from person identity PINs using facial and voice information [83], to generative feature learning for image and text [33]. Text-to-image retrieval which is a subcategory of cross-modal matching is a well-studied problem in computer vision facilitated by datasets describing birds, flowers or regular objects [92, 134, 71]. However, person search using textual descriptions is a relatively new

application that emerged in 2017 when the CUHK-PEDES dataset [64] was published. CUHK-PEDES contains images of individuals along with two textual descriptions for each image. The progress in the past two years on this benchmark has been remarkable with the rank-1 accuracy increasing from 20% to 50%. Most methods [64, 63] rely on a mostly similar procedure: (i) extract discriminative image features using a deep neural network, (ii) extract text features using an LSTM, and (iii) propose a loss function that measures as accurately as possible the distance between the two features. To improve the performance, Chen *et al.* [11] proposed two loss functions that aimed to perform image-language association at a global level (whole sentence to whole image) as well as at a local level (phrases in the sentence with image parts extracted from attention blocks). Zhang and Lu [138] followed a different approach by treating person search as a matching problem in which image features are projected into the text domain and vice-versa and KL divergence losses are utilized to learn discriminative feature representations.

Image-based Person Search: Image-to-image person search is the retrieval of relevant whole-scene images of an individual given a probe image without relying on manually cropped images of pedestrians. Instead of trying to solve separately the tasks of pedestrian detection and person re-identification, image-based person search methods jointly solve both problems in a single framework [128]. Some works [128, 127] rely on sub-networks that propose pedestrian regions that are then used extract features for human re-identification. In the work of Xiao *et al.* [128] region proposals are first produced (similar to the way R-CNN produces region proposals) which are then fed to an identification network for feature extraction. An online instance matching loss is then introduced that maintains labeled identity proposals in a look-up-table and unlabeled identity proposals into a circular queue. Other works focus on extracting discriminative features at different spatial resolutions [59] or leverage temporal information when video information is available [42].

Chapter 3

Objective 1: Define an attribute ontology

In this chapter, the objective is to develop an ontology for the identity and the identity-related attributes. In order to perform text-to-image retrieval in a person search application, focusing on accurately measuring the distance between the features of the two modalities is insufficient. Aiming to introduce structure to the free-form textual input, an attribute ontology is introduced. Word tokenization and part-of-speech tagging are performed to the input sentence to extract all the nouns and adjectives that describe the depicted individual. These extracted traits are then mapped to the attribute ontology that generates positive or negative labels for the set of attributes that it includes. For example, if the description contains words such as “man”, “guy”, “boy” then the *“Sex”* attribute of the ontology has a positive label. Using this ontology extract attribute pseudo-labels are extracted that can then be used to train attribute classification models. This process requires no additional supervision as only the attribute ontology (and the mapping) need to be constructed once before training starts. By using the proposed ontology and leveraging attribute classification as an auxiliary task, our model can learn better feature representations. The primary contribution is a method with the following lessons learned:

- Learning to match textual to image features and vice versa in an end-to-end learning framework is of paramount importance. The baseline results indicate that jointly solving the problems of person identification and cross-modal matching can achieve superior results than the best existing performing method.

- Introducing structure to the textual input through the proposed attribute ontology can be beneficial. By extracting pseudo-visual attribute labels from the textual descriptions and adding attribute classification as an auxiliary task to our network performance improvements are observed without any additional need for data annotation.

- The proposed method can be applied to out-of-distribution textual queries originating from the University of Houston police department with satisfactory retrieval results.

3.1 Method

In this work, the proposed method named is introduced which named *TIPS*: a Text-to-Image Person Search approach that effectively retrieves the most relevant images of humans given a textual description as an input. During training, our aim is to learn discriminative visual and textual feature representations capable of accurately retrieving the ID of an individual. The training procedure is depicted in Figure 3.1 and is described in detail below. Specifically, the input at training-time consists of triplets (X_i, T_i, Y_i) where X_i is the image of the human, T_i, is the textual description describing that image, and Y_i is the identity of this human. To learn the visual representations denoted by $\phi(X_i)$ a ResNet-101 network is used as a backbone network. The feature map of the last residual block is projected to the dimensionality of the feature vector using a global average pooling and a fully-connected layer. For the textual input, each word is represented as a D-dimensional

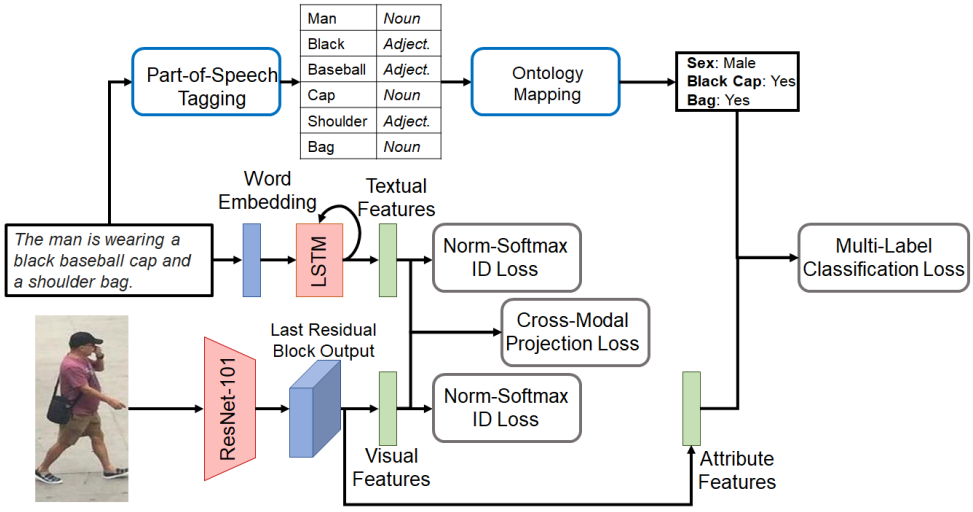

Figure 3.1: Given an image of a person along with the corresponding textual description discriminative visual and textual representations are learned to perform effective person search. Textual and visual features are learned through their corresponding sub-networks which are then fed to an identification loss as well as a cross-modal projection loss function.

one-hot vector where D is the vocabulary size. Each one-hot vector is mapped to a word embedding and is fed to a long short-term memory network (LSTM) which effectively summarizes the content of the input textual description. Finally, the textual representation denoted by $\tau(T_i)$ is obtained by projecting the output of the LSTM to the dimensionality of the feature vector using a fully-connected layer. Following the work of Zhang and Lu [138], two separate loss functions are employed to train our baseline network: (i) a norm-softmax cross entropy loss for identification and (ii) a cross-modal projection matching loss. The norm-softmax cross entropy loss [72, 120] introduces an L2-normalization on the weights of the output layer. By doing so, it enforces the model to focus on the angle between the weights of different samples to perform identification instead of their magnitude. For the visual features, the norm-softmax cross entropy loss can be described as

follows:

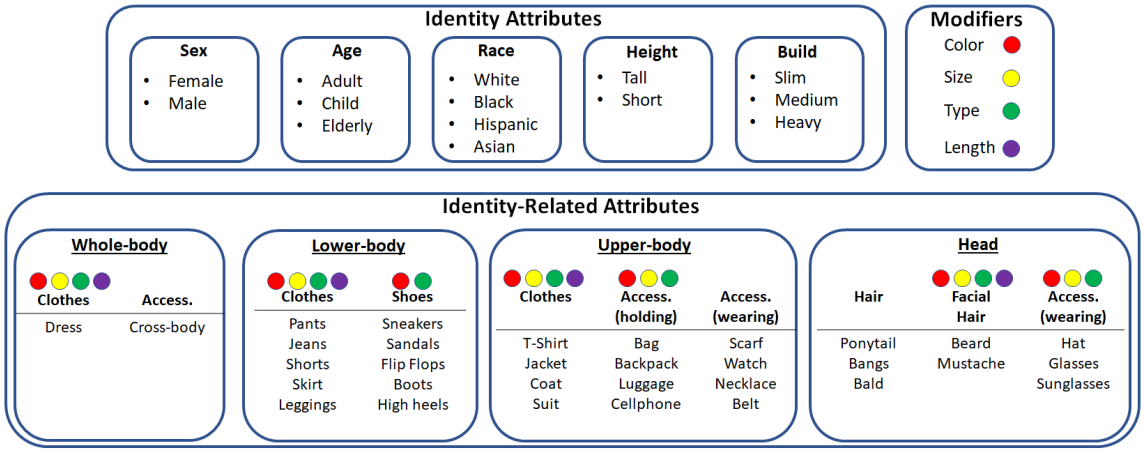

Figure 3.2: Our proposed ontology groups the visual attributes to identity and identity-related based on whether they can change over a short period of time or not. The identity-related attributes can be further split into fine-grained categories based on which part of the body they belong to. Four modifiers are also provided that can describe different attributes of the ontology.

$$L_s^v = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{\exp(W_i^T\phi(X_i)+b_i)}{\sum_j \exp(W_j^T\phi(X_i)+b_j)}\right), s.t. ||W_j|| = 1\ , \tag{3.1}$$

where W_i, b_i are the weights and the bias of the classification layer for the visual feature representation $\phi(X_i)$. The loss for the textual features L_s^t is computed in a similar manner. However focusing solely at performing accurate identification is not sufficient since the goal is to perform cross-modal retrieval at test time. Towards this direction, the cross-modal projection-matching loss [138] is used which incorporates the cross-modal projection into KL divergence to associate the representations across different modalities. Intuitively, the larger the scalar projection from one modality to another, the more similar the two representations are. The text representation is first

normalized $\bar{\tau}(T_j) = \frac{\tau(T_j)}{||\tau(T_j)||}$ and then the probability of matching $\phi(X_i)$ to $\bar{\tau}(T_j)$ is:

$$p_{i,j} = \frac{\exp\left(\phi(X_i)^T \bar{\tau}(T_j)\right)}{\sum_{k=1}^{B} \exp\left(\phi(X_i)^T \bar{\tau}(T_k)\right)} , \tag{3.2}$$

where B is the batch size. Since in each mini-batch there might be more than one positive matches (i.e. visual and textual features originating from the same identity) the true matching probability is normalized as follows:

$$q_{i,j} = \frac{Y_{i,j}}{\sum_{k=1}^{B} Y_{i,k}} . \tag{3.3}$$

The cross-modal projection matching loss of associating $\phi(X_i)$ with correctly matched text features is then defined as the KL divergence from the true matching distribution q_i to the probability of matching p_i. For each batch this loss is defined as:

$$L_v = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \epsilon}\right), \tag{3.4}$$

where ϵ is a very small number. The same procedure is followed to perform the opposite matching (i.e. from text to image to compute loss L_t) and the summation of the two individual losses constitutes the cross-modal projection matching loss $L_c = L_v + L_t$.

3.2 From Textual Descriptions to an Ontology

Learning textual and visual features in a joint manner, using the norm-softmax and cross-modal projection matching losses, addresses the challenge of accurately measuring the similarity between the two feature vectors. A major limitation of the existing approaches is the simplicity in which the textual representation is handled. Unlike the image input which is fed to a deep architecture that extracts discriminative visual features, the textual features are learned using just a bidirectional LSTM. The input sentences contain typos and mistakes, words are written in American as well as

British English (e.g., gray and grey) and sentences demonstrate a large variance in what is described even for the same image.

To overcome this limitation one could (i) identify the most frequently used words in the training set (e.g., man, woman, shirt, bag, pants), (ii) treat them as attribute classes that are positive whenever such words appear in the textual description, and (iii) train an attribute classifier that learns to predict such attributes given the image features. Such an approach is a step in the right direction since it adds structure to the text input and tries to learn features that are better for text-to-image retrieval. However, choosing which words to use and how many as attributes is subjective and it is not clear which attribute labels will lead to better performance versus some others. Additionally, solely focusing on words is insufficient as attributes related to colors can describe multiple things at the same time.

Thus, an attribute ontology is proposed that maps textual description to attribute classes. Given a textual description, word tokenization part-of-speech tagging is performed to extract all nouns and adjectives. The extracted nouns and adjectives are then mapped to an ontology that aspires to introduce structure to the free-form text and help the network learn better feature representations.

3.2.1 Attribute Ontology

To better understand the visual attributes included in the textual descriptions a distinction is made between identity and identity-related attributes. The former are attributes that cannot easily change between subsequent days (e.g., sex, age, race, height, and build) whereas the latter are attributes such as clothes or accessories that an individual can change over a short period of time. The proposed ontology is depicted in Figure 3.2. Attributes are split based on the body-parts they correspond to (head, upper-body, lower-body, full-body). Such information can be leveraged if

the 2D pose is predicted within the network. For the identity-related attributes, a few *"modifiers"* are also introduced which are attributes (usually adjectives) that are used to better describe such traits. By observing the training textual description the following modifiers were identified: (i) the color, (ii) the size, (iii) the length, (iv) the type. For example, a "blue, V-shaped, large t-shirt" describes a t-shirt with modifiers color, type, and size. Datasets such as the PETA benchmark [20] have separate classes for each of the modifiers given an attribute (e.g., black t-shirt is a separate class from a gray t-shirt). Although this makes attribute classification systems easier to build, it introduces a large class imbalance (gray t-shirts are way less likely to appear in a dataset compared to t-shirts which might already be imbalanced as a class). The the same time such an approach is limited since extracting all possible configurations of $\{$color, type, size, length$\}$ for each individual attribute extracted from textual descriptions is intractable.

3.2.2 Visual Attribute Classification

Once the input text is mapped to the ontology, attribute labels can be extracted to perform visual attribute classification. All attributes are treated as binary labels (i.e. whether that attribute exists or not). The last residual block of our backbone architecture is fed to a convolutional layer with a 1×1 kernel followed by a fully-connected layer which maps the attribute-related features to attribute classes.

To train our network a modified weighted binary cross entropy loss is used:

$$L_B^w = -\frac{1}{N}\frac{1}{C}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c a_{i,c}\log(\hat{a}_{i,c}) + (1-a_{i,c})\log(1-\hat{a}_{i,c}), \tag{3.5}$$

where C the number of the attribute classes, $a_{i,c}, \hat{a}_{i,c}$ are the ground-truth and prediction pairs for each attribute respectively, and $w_c = \exp(-a_c)$ is the weight for c-th attribute. Note that a_c is the prior distribution of the c-th attribute in the training set. This cost-sensitive learning

approach is used to account for the attribute imbalance. Alternative losses such as the focal loss were investigated to tackle class imbalance but demonstrated inferior performance.

3.2.3 Training and Testing Details

The loss function that is used to train our network is the summation of the four individual losses:

$$L = L_s^v + L_s^t + L_c + L_B^w. \tag{3.6}$$

Stochastic gradient descent (SGD) was used with momentum equal to 0.9 to train the image and attribute networks and the Adam optimizer [54] for the textual networks. The learning rate was set to 2×10^{-4} and was divided by ten when there the rank-1 performance plateaued at the validation set until 2×10^{-6}. The batch-size was set to 64 and the weight decay to 4×10^{-4}. The dimensionality of all feature vectors was set to 512. Regarding the ontology, it was found that by mapping textual descriptions to 11 distinct classes (depicted in Table 3.2) a significant performance improvement can be obtained without adding a computational overhead. The modifier information was not leveraged to avoid learning very specific attribute classifiers with highly imbalanced data (e.g., blue v-neck t-shirts) as these would degrade our performance.

At testing time given a textual description, its textual features ($\tau(T_i)$) are extracted and their distance between all image features ($\phi(X_j)$) in the test set is computed using the cosine similarity:

$$\cos(\theta) = \frac{\tau(T_i)\, \phi(X_j)}{||\tau(T_i)||\, ||\phi(X_j)||}. \tag{3.7}$$

The distances are then sorted and rank-1 through rank-10 results are reported.

3.3 Experiments

3.3.1 Dataset Description and Evaluation Metrics

To evaluate our method the CUHK-PEDES [64] dataset was used which is the only publicly available dataset for person search using textual descriptions. It consists of 40,206 images of individuals of 13,003 identities, and each image is described by two textual descriptions. The dataset is split into 11,003/1,000/1,000 identities for the training/validation/testing sets with 34,054, 3,078 and 3,074 images respectively in each subset. All images are resized to 256×128. Following the data pre-processing steps of Li *et al.* [64] textual descriptions longer than 50 words are trimmed and words that appear less than three times in the whole training set are discarded before creating the vocabulary. Since person search is a retrieval problem, the most relevant results are retrieved given a query textual description and report rank-1, rank-5 and rank-10 results for each method.

3.3.2 Quantitative Results

Our approach is evaluated against all 12 methods that have been tested on the CUHK-PEDES dataset. Some key methods that have been evaluated on this dataset include (i) the method of Li *et al.* [63] which learns discriminative features using two attention modules working on the both modalities at different levels but it is not end-to-end; (ii) the work of Chen *et al.* [11] which identify local textual phrases and tries to find the corresponding image regions using an attention mechanism; (iii) and the method of Zhang and Lu [138] in which two projection losses are proposed to learn features for text-to-image matching. The obtained results are presented in Table 3.1 which demonstrates that our proposed approach achieves state-of-the-art results by improving the rank-1 accuracy by 2.33%. Our improvements over previous works (the relative improvement is 4.72%)

Table 3.1: Text-to-image retrieval results (%) on the CUHK-PEDES dataset. The results are ranked based on the rank-1 accuracy.

Method	**Rank-1**	**Rank-5**	**Rank-10**
iBOWIMG [141]	8.00	-	30.56
Word CNN-RNN [92]	10.48	-	36.66
Neural Talk [118]	13.66	-	41.72
GMM+HGLMM [55]	15.03	-	42.47
deeper LSTM Q+norm I [3]	17.19	-	57.82
GNA-RNN [64]	19.05	-	53.64
IATV [63]	25.94	-	60.48
PWM-ATH [13]	27.14	49.45	61.02
GLA [11]	43.58	66.93	76.26
Dual Path [140]	44.40	66.26	75.07
CMPM + CMPC [138]	49.37	-	79.27
TIPS	**51.70**	**74.33**	**82.39**

are originating from the proposed attribute ontology that introduces structure to the textual input and extracts attribute pseudo-labels which are then used to improve the retrieval results. This indicates that by combining the joint feature learning with the auxiliary task of attribute classification the network learns more discriminative visual and textual features.

During training, it was also observed that for the visual attribute classification task, the mean average precision (mAP) on the validation set increased from 40% in the first epoch to 68%. This indicates that besides learning features capable of performing image retrieval our method can also

learn attribute-specific features that are used for classification. In Table 3.2 mean average precision results (mAP) are reported for the 11 most prevalent attribute classes when three different loss functions are used for training. The first loss is the standard binary cross-entropy loss, the second is the weighted binary cross-entropy described in Eq. 3.5 and the third is the weighted focal loss described in [70, 98]. Both of the last two loss functions handle class imbalance at a class-level by introducing weight-specific class weights learned from the training set, whereas the weighted focal loss focuses also at an instance-level. This is done by weighing the contribution of each sample to the final loss based on how far is the probability prediction from the original label (raised to a power γ). Attributes that are fairly balanced such as sex, shoes, and t-shirt have a high AP compared to imbalanced attributes such as jacket or glasses that have a poor AP.

Table 3.2: Mean average precision results (mAP) on the CUHK-PEDES dataset using as ground-truth the pseudo-labels extracted from the textual descriptions using the proposed attribute ontology.

Attribute	**Cl. Imbalance**	$\mathcal{L}_B$	$\mathcal{L}_B^w$	$\mathcal{L}_F$
Sex	1:1	94.22	94.11	**94.56**
T-shirt	1:2	89.03	**89.05**	88.95
Jacket	1:8	46.99	**50.69**	48.56
Dress	1:17	50.38	**61.01**	53.95
Pants	1:3	75.42	**75.63**	74.65
Backpack	1:11	59.96	**57.79**	53.49
Glasses	1:13	31.27	**41.22**	40.69
Shoes	1:1	73.11	73.42	**74.28**
mAP	-	65.42	**68.25**	66.92

Table 3.3: Ablation studies on the CUHK-PEDES dataset to asses the impact of individual modules on the final performance of our method.

Method	**Rank-1**	**Rank-5**	**Rank-10**
Baseline w/ ResNet-50	46.05	69.05	78.67
Baseline w/ ResNet-101	49.85	72.94	80.48
Baseline w/ Attributes	50.14	73.50	81.96
Baseline w/ Ontology	**51.70**	**74.33**	**82.39**

3.3.3 Ablation Study

Aiming to obtain a better understanding of the contributions of each individual component towards the final performance an ablation study is conducted. To investigate to what extent the primary network affects the final performance a ResNet-50 backbone architecture is first employed and then its depth is increased to 101. This is because it is commonplace that as architectures become deeper, the impact of individual add-on modules becomes less significant. By comparing the first two lines of results of Table 3.3 it becomes apparent that the increase of depth improves the rank-1 accuracy by 3.8%. Note that in both cases the networks were trained to perform identification and feature matching and no attribute ontology was utilized. In addition, our baseline architecture achieves better accuracy than the previous state-of-the-art [138]. This is because: (i) a bigger backbone is used for the image input and (ii) the output of the LSTM was projected to a fully-connected layer and learn its weights instead of simply performing max-pooling. In the next experiment, the top-10 most frequently used words were identified and by treating them as attribute classes (without any ontology mapping) an attribute classifier is trained to learn to predict such attributes. This is process is performed at the same time with the feature learning for person search in an end-to-end manner.

The results are improved compared to the original baseline but the increase is very small in terms of rank-1 accuracy. Focusing on some meaningful attribute classes related to identity attributes such as sex and identity-related attributes such as clothing and accessories and mapping the textual input to the proposed ontology can boost the performance and result in a better image retrieval system.

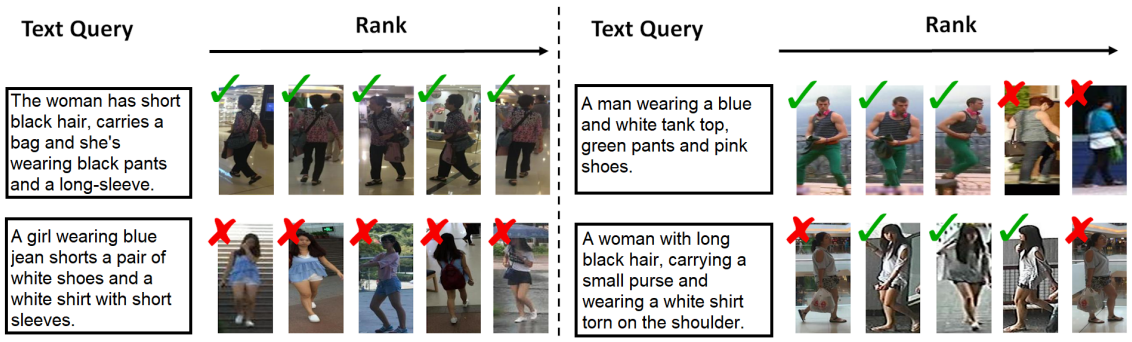

Figure 3.3: Qualitative results on the CUHK-PEDES dataset. Given a textual description as a query, the top-5 most relevant images are ranked from left to right. Successful image retrieval is performed in cases with poor lighting, under different poses, and with different visual attributes. Even in failure cases, the retrieved results are still very relevant (e.g., in the bottom left example the color of the shirt and pants are the opposite from the input text).

3.3.4 Qualitative Results

To evaluate the performance of the proposed approach two sets of qualitative results are provided. The retrieved images are presented in Figure 3.3. Our method is capable of learning cloth and accessory-related correspondences as it can accurately retrieve images of people carrying bags with the correct set of clothing. For example, in the bottom right query even the two incorrectly retrieved results contain images of females that match the textual description even if the identity

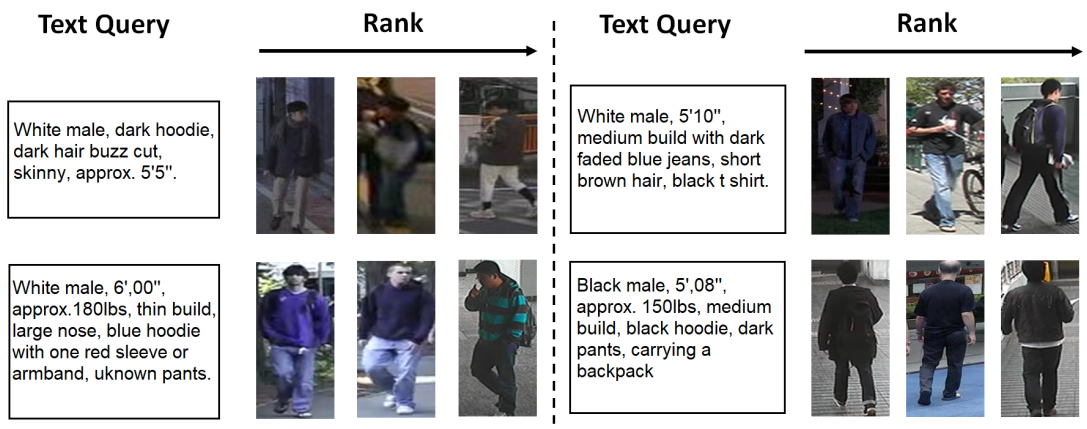

Figure 3.4: Qualitative results on the CUHK-PEDES dataset when the input textual query is from the University of Houston police department suspect description list. Our method successfully retrieves humans with blue hoodies and pants on the bottom left and individuals with dark-colored clothes on the bottom right.

of that individual is not the right one. Even in the bottom left example in which all top-5 retrieved images contain a different identity our method retrieves images of females with the correct set of clothes (shoes, shorts, and shirt) and the correct colors for the top two images but with the colors of the shirt and shorts retrieved the other way around.

Aiming to evaluate how well our method performs when the input textual descriptions are originating from a different distribution (and not from the descriptions of the CUHK-PEDES dataset) the suspect descriptions from the University of Houston police department (UHPD) which were publicly available between the years 2013 and 2018 were collected. After mapping them to the existing vocabulary to perform word tokenization these descriptions were used as text queries at test-time to qualitatively investigate how successfully our method retrieves relevant images. In Figure 3.4 the top-3 retrieved images from the CUHK-PEDES dataset are depicted when the input textual description is from UHPD descriptions. In most cases, the top-5 retrieved results contain

at least two times the same identity which shows the consistency of our method in retrieving relevant images. For example, in the bottom left query all three individuals wear blue hoodies, carry backpacks and wear long pants. In the bottom right textual description, all three results have dark pants and black upper body clothes. These results are still far from perfect and there is room for improvement since attributes such as “polo shirt” are not retrieved at all in the upper right textual input.

Chapter 4

Objective 2: Develop an algorithm for extracting identity attributes from images

In this chapter, two algorithms that extract soft-biometrics using privileged information are first presented. Then two additional algorithms that extract visual attributes using neural networks are provided.

4.1 Gender Classification using Human Metrology

4.1.1 Gender Prediction with LUPI

Below, a detailed review of the concepts of SVM is provided for completeness, and the SVM+ algorithm as well as the Margin Transfer method of Sharmanska *et al.*[102] are provided. Throughout this section the same notation as in the work of Sharmanska *et al.*[102] is used.

Ratios of anthropometric measurements: Based on the findings of the work of Cao *et al.*[7], using the actual values of anthropometric measurements (e.g., limb lengths in *mm*) from an anthropometric database results in good gender classification accuracy. However, such information cannot be accurately obtained from state-of-the-art computer vision algorithms without employing depth information (e.g., use data obtained from a Kinect RGB-D sensor). To address this limitation, in this work we propose to use ratios of anthropometric measurements. Hence, errors during the estimation of the actual values will be alleviated. A variety of anthropometric measurements from the body and the head is provided in the CAESAR database [94]. These measurements are split into two groups. The first group contains only ratios of body measurements that can be captured from a regular surveillance camera and computed from state-of-the-art computer vision algorithms. This set, which will be denoted as X, contains only observable information (e.g., arm or leg lengths) and will be available during both the training and the testing phases. The second group, which will be denoted by X^*, contains ratios of body measurements that are difficult to obtain with an automated acquisition system (e.g., circumferences of body parts) as well as a few measurements that correspond to the head (e.g., head breadth or face length). This type of information is considered as privileged and it will not be available at test time.

From SVM to SVM+: In the standard paradigm of supervised learning for binary classification, the training set consists of N pairs of feature vectors x_i, along with their respective labels y_i, represented as $(x_1, y_1), \ldots, (x_N, y_N)$, $x_i \in \mathbb{R}^d$ where d is the number of features of each sample and $y_i \in \{-1, +1\}$. The standard SVM classifier finds a maximum-margin separating hyperplane between the two classes. In a LUPI setup, during the training phase, sets of triplets $(x_i, x_i^*, y_i), x \in \mathbb{R}^d, x^* \in \mathbb{R}^{d^*}$, $y_i \in \{-1, +1\}$ are provided, where feature vectors x^* represent the additional (i.e. privileged) information. During the testing phase, features from the privileged space X^* are not available. The goal of LUPI is to exploit the privileged information during the training phase to learn a model that further constrains the solution in the original space X and thus, it can more

accurately describe the testing data. In this paradigm, the slack variables ξ_i are parameterized as a linear function of privileged information $\xi_i(w^*, b^*) = \langle w^*, x_i^* \rangle + b^*$. The SVM+ problem in the training phase solves the following minimization problem:

$$\begin{aligned} \underset{w,\, b,\, w^*, b^*}{\text{minimize}} \;\; & \frac{1}{2}\big(||w||^2 + \gamma||w^*||^2\big) + C\sum_{i=1}^{N} \xi_i(w^*, b^*). \\ \text{subject to:} \;\; & y_i\big(\langle w, x_i \rangle + b\big) \geq 1 - \xi_i(w^*, b^*) \\ & \xi_i(w^*, b^*) \geq 0,\; i = 1, \ldots, N\ . \end{aligned} \tag{4.1}$$

Margin Transfer: Sharmanska *et al.*[102] investigated the framework of using privileged information for object recognition and introduced a Margin Transfer approach. The proposed method interprets the LUPI concept as learning the *easiness* and *hardness* of each sample to be classified based on the margin distance to the classifying hyperplane in the privileged space. This knowledge is then transferred to the original space to train a classifier with improved performance. A standard SVM classifier is first trained on X^*, the prediction function of which is $f^*(x^*) = \langle w^*, x^* \rangle$, and the margin distance $\rho_i := y_i f^*(x_i^*)$ between the training samples and the decision function in the privileged space is computed. Large values of ρ_i indicate that the respective sample can be classified easily, low values correspond to samples that are more difficult to classify, and negative values samples that are impossible to classify. The minimization problem is formulated as follows:

$$\begin{aligned} \underset{w \in \mathbb{R}^d,\, \xi_i \in \mathbb{R}}{\text{minimize}} \; & \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \xi_i. \\ \text{subject to:} \; & y_i\langle w, x_i \rangle \geq \rho_i - \xi_i \\ & \xi_i \geq 0,\; i = 1, \ldots, N\ . \end{aligned} \tag{4.2}$$

Unlike SVM+, the performance of the classifier in the privileged space is very important for the Margin Transfer method because the information in the privileged space defines the margin in the original space.

Table 4.1: Gender classification mean accuracy (%) and standard deviation on the CAESAR dataset using SVM and SVM+.

Testing Features	SVM	SVM+
X	97.61 ± 0.44	$\mathbf{98.18 \pm 0.56}$
X_L	95.34 ± 0.74	$\mathbf{95.82 \pm 0.81}$
X_U	$\mathbf{76.69 \pm 2.98}$	76.54 ± 2.95
$X \cup X^*$	99.10 ± 0.23	-
Cao *et al.*[7]	99.37	-

4.1.2 Experimental Evaluation

The CAESAR database [94] is employed which is comprised of 44 anthropometric measurements (in *mm*), the weight (in *kg*) and the gender of 2,392 US and Canadian civilians. After data-preprocessing and discarding data with missing values, the size of the dataset used for the experimental evaluation is 2,369$\times$39 including the gender. The ratios of anthropometric measurements are split to: (i) X which contains ${}^{(12\times11)}/_{2} = 66$ observable features (i.e. ratios) for each human subject and (ii) the privileged set X^* with size of ${}^{(26\times25)}/_{2} = 325$ for each sample. When no privileged information is used, a linear SVM is utilized which requires only the penalty parameter C to be cross-validated. For SVM+, a linear kernel was used in the original space and a radial basis function kernel type in the correction space. In the latter case, additional tuning of the kernel coefficient γ in the correcting space is necessary. The search-space for both C and γ were $[10^{-4}, 10^{-3}, \ldots, 10^{4}]$. A standard 5-fold cross-validation scheme was selected and a full-grid search was performed.

Gender Classification on the CAESAR dataset: In Table 4.1, we present the results of the proposed approach using the standard SVM and SVM+ methods when the testing set comprises observable features: (i) from the whole human body (X), (ii) only the lower body (X_L), and (iii) only the upper body (X_U). The first column denotes which of the observable features are available at test time. The last two rows employ all the available information during both training and testing, and thus a LUPI framework is not applicable. The last two rows correspond to classification results when all the measurements are used in both training and testing and can be interpreted as an upper boundary for the classification performance. The second to last row uses ratios of anthropometric measurements as features, whereas the method of Cao *et al.*[7] uses the corresponding actual values of the measurements. When all the observable features are used, the LUPI paradigm improves gender classification accuracy. Interestingly, features from the lower part of the body exhibited a significantly better classification accuracy compared to the ones obtained from the upper body. When only upper body features are available at test time, the LUPI framework does not appear to result in increased accuracy, and in both methods the classification accuracy is reduced to $76.6\,\%$ while the standard deviation increases to almost $3\,\%$.

Leveraging Privileged Information in Human Metrology: Beneficial or Redundant? Based on the results obtained from the SVM+ method, the LUPI framework can improve the classification accuracy. However, two important characteristics of LUPI have to be taken into consideration. The first is what may be considered privileged information. Although conceptually it might be appealing to use circumferences of human limbs as prior information to boost the gender prediction accuracy, this is not always true as demonstrated by the experimental results (when only upper body features are visible). The second is how the existence of privileged information is exploited to improve the accuracy during test time.

Gender classification results are reported in Table 4.2 using a standard SVM for 20 random

Table 4.2: Gender classification accuracy (%) on still images from the PaSC and SARC3D datasets using an SVM when the feature set contains full, upper and lower-body features.

	Set of features		
Dataset	X	X_L	X_U
PaSC	71.37 ± 1.64	57.65 ± 2.82	58.06 ± 2.73
SARC3D	86.00 ± 2.00	78.00 ± 4.00	72.00 ± 4.00

train-test splits. The reason the reported classification accuracy on SARC3D is higher than PaSC, can be attributed to the fact that the 3D poses in SARC3D are easier to be estimated since people are standing in an upright position. Columns two to four in Table 4.2 contain different results depending on which parts of the human body are visible. Identifying the gender of humans from real images using anthropometric measurements is a challenging task. Its difficulty arises from the fact that the 3D pose estimation algorithm starts with an initial 3D pose and a dictionary of poses (i.e. bases) and by exploiting this information maps the 2D joint locations to 3D through an optimization scheme. The performance of this algorithm is sensitive to the initialization, and thus the performance is not always robust. Note that employing privileged information from the CAESAR dataset, and exploiting this information at test time using as features the obtained measurements from the images, resulted in a worse performance. Thus, the LUPI paradigm using anthropometric measurements estimated from images was not investigated further.

4.2 Predicting Privileged Information for Height Estimation

4.2.1 Privileged Information for Height Estimation

ϵ-SVR+: The SVM+ method was proposed by Vapnik and Vashist [116] to exploit privileged information for binary classification tasks. They also generalized the LUPI paradigm for the regression estimation task denoted by ϵ-SVR. The goal in support vector regression is to find a function that has at most ϵ deviation from the obtained targets y_i for the training set and is as flat as possible [105]. This means that as long as the errors are less than ϵ they are not taken into consideration. However, any deviation larger than this will not be accepted. The standard soft-margin ϵ-SVR is formulated by the following optimization problem:

where $w \in \mathbb{R}^m$ represents the weight vector, $||w||^2$ indicates the size of the soft margin and $b \in \mathbb{R}$ is the bias parameter. Additionally, ξ_i is the slack variable for one training sample and indicates the deviation from the margin borders and C denotes the penalty parameter. Note that the ξ^* in ϵ-SVR has nothing to do with the privileged space. It denotes the space of width ϵ below the margin as depicted in a toy example in Figure 4.1. When privileged information is available at training time, three sets of linear functions are considered. The first set lies in the observable space in which the decision function is approximated while the other two are functions that approximate the correcting functions for the slack variables ξ_i and ξ_i^*. The optimization problem is formulated

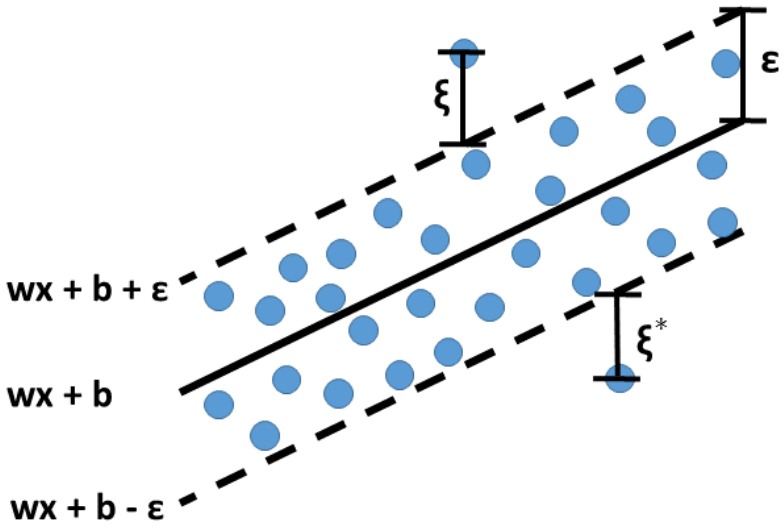

Figure 4.1: The ϵ-insensitive band for a one-dimensional linear support vector regression problem.

as:

$$\begin{aligned}
&\underset{\substack{w,\, w_1^*,\, w_2^* \\ b,\, b_1^*,\, b_2^*}}{\text{minimize}} \quad \tfrac{1}{2}\big(||w||^2 + \gamma(||w_1^*||^2 + ||w_2^*||^2)\big) + \\
&+ \quad C\sum_{i=1}^{l}(\langle w_1^*, x_i^*\rangle + b_1^*) + C\sum_{i=1}^{l}(\langle w_2^*, x_i^*\rangle + b_2^*)
\end{aligned}$$

$$\begin{aligned}
\text{subject to:} \quad y_i - \langle w, x_i\rangle - b &\leq \epsilon + \langle w_1^*, x_i^*\rangle + b_1^* \\
\langle w, x_i\rangle + b - y_i &\leq \epsilon + \langle w_2^*, x_i^*\rangle + b_2^* \\
\langle w_1^*, x_i^*\rangle + b_1^* &\geq 0 \\
\langle w_2^*, x_i^*\rangle + b_2^* &\geq 0 \\
i = 1 \ldots l\ .&
\end{aligned} \tag{4.3}$$

where parameters with sub-indices equal to one and two correspond to the first and second correcting functions, respectively.

Privileged Information Prediction (PIP): A novel method of estimating the height using support

Table 4.3: Height estimation error (%) for a regular ϵ-SVR, ϵ-SVR+ and the proposed privileged information prediction (PIP) approach.

	ϵ-SVR+			**PIP**		
Q	Male	Female	Both	Male	Female	Both
1^{st}	3.95 ± 0.34	4.17 ± 0.27	4.28 ± 0.33	4.31 ± 0.24	4.27 ± 0.22	$\mathbf{3.96 \pm 0.34}$
2^{nd}	1.68 ± 0.21	1.77 ± 0.11	$\mathbf{2.50 \pm 0.16}$	1.80 ± 0.12	1.71 ± 0.19	2.65 ± 0.12
3^{rd}	1.87 ± 0.17	1.84 ± 0.11	2.71 ± 0.19	1.84 ± 0.15	1.66 ± 0.13	$\mathbf{2.69 \pm 0.11}$
4^{th}	4.30 ± 0.26	4.01 ± 0.25	3.86 ± 0.33	3.96 ± 0.26	3.97 ± 0.30	$\mathbf{3.73 \pm 0.22}$
All	$\mathbf{2.94 \pm 0.13}$	2.91 ± 0.12	3.33 ± 0.10	2.95 ± 0.12	$\mathbf{2.89 \pm 0.10}$	$\mathbf{3.25 \pm 0.12}$

vector regression (ϵ-SVR) in two steps is proposed. Our method takes as an input the two groups of observable and privileged human measurements and outputs its height, which is then mapped to classes (i.e. quartiles) that correspond to percentile ranges. For the purposes of this work, it is assumed that the anthropometric measurements of a human are available and provided to the system. However, in a real-life scenario, the observable measurements are obtained from an image of a human by applying a 3D pose estimation algorithm to obtain the location of the joints in three dimensions. The estimated skeleton is used to derive the observable measurements (e.g., arm length, hip to knee length). Another group of features is utilized (i.e. privileged measurements) such as circumferences of body parts which will be available during the training phase. Ratios of anthropometric measurements are computed for each of them, to alleviate the error that would occur during the estimation of the actual values. The privileged vector $\boldsymbol{x}^*$ is then used to find the K most informative features denoted by $\hat{\boldsymbol{x}}^*$ using the minimum redundancy maximum relevance

Algorithm 4.1: Privileged Information Prediction (PIP)

Input : Ratios of observable $\boldsymbol{x}$ and selected privileged $\boldsymbol{x}^*$ features , labels $\boldsymbol{y}$, number of selected features K, ϵ, estimation error allowed e

for *i= 1,...,K* **do**

  // privileged feature prediction

  $\hat{\boldsymbol{x}}_i^* \leftarrow \epsilon$-SVR model trained on $(\boldsymbol{x}, \boldsymbol{x}_i^*)$

end

// height estimation

$h \leftarrow \epsilon$-SVR model trained on $([\boldsymbol{x}^T \ \hat{\boldsymbol{x}}^{*T}]^T, \boldsymbol{y})$

$h_c \leftarrow$ mapping to height classes by allowing error e

Output: Height h in *cm*, $h_c \in \{1^{st}, 2^{nd}, 3^{rd}, 4^{th}\}$ quartiles

feature selection (mRMR) of Peng *et al.* [87] with the mutual information difference (MID) feature selection scheme. For each selected feature, a support vector regression model is learned from $\boldsymbol{x}$ that predicts its value $\hat{\boldsymbol{x}}_i^*$. A new feature vector is formed which contains the concatenation of $\boldsymbol{x}$ with the K predicted values of $\hat{\boldsymbol{x}}^*$ and a new regression model is trained to predict the height. Since height is a continuous variable, performing classification (i.e. 1^{st} quartile) would imply that the boundaries of the classes would have to be strictly defined, which would result in many misclassification errors. To address this challenge, a percentage of error is allowed between the predicted height value and the actual height (i.e. ground truth value). Thus, if a testing sample is misclassified but the error (%) in the estimation from the actual value is less than a threshold then this sample is considered as correctly classified. Classification accuracy results are reported in Section 4.2.2.

The key differences between ϵ-SVR+ and the proposed approach are that the former uses information from the privileged space to add an extra term to the optimization function and further constrain the solution in the observable space. In contrast, our method employs the predictions of

privileged features as extra information that can be used to estimate the height. This implies that at testing time the proposed method contains an estimation error in the feature vector that is used to predict the height. This is not the case for the ϵ-SVR+ algorithm. Unlike ϵ-SVR+, the proposed approach, is significantly faster to train and cross-validate despite the two regression steps instead of one. That is because the parameters that have to be tuned, except the parameter ϵ, are two for a Gaussian kernel instead of four. Finally, to the best of our knowledge, an implementation of ϵ-SVR+ is not currently available, whereas the proposed approach can be re-implemented using standard programming packages.

4.2.2 Experimental Evaluation

For the purposes of this work, the CAESAR database [94] was used which comprises 44 anthropometric measurements (in *mm*) such as the spine-to-elbow length or the chest circumference, the weight (in *kg*), and the gender of 2,392 US and Canadian civilians. After data preprocessing and discarding data with missing values, the size of the dataset for the experimental evaluation is 2,369 with 39 features for each sample, including the gender. The number of observable and privileged features are 11 and 27, respectively, whereas gender is investigated separately. Thus, the ratios of anthropometric measurements obtained are split into: (i) $\boldsymbol{x}$ which contains $^{11\times10}/_{2} = 55$ observable features (i.e. ratios) for each human subject and (ii) the privileged $\boldsymbol{x}^*$ with size of $^{27\times26}/_{2} = 325$ for each sample. A Gaussian kernel was used for all three methods (i.e. SVR, SVR+, and PIP), which means that besides the cost parameter C, the width γ_G of the kernel needs to be cross-validated. In the case of SVR+, there is an additional γ_G that needs to be cross-validated along with the parameter γ of the correcting space as shown in Equation 4.3. The possible values for all parameters were $[10^{-4}, 10^{-3}, \ldots, 10^{4}]$, and a standard 5-fold cross-validation scheme was employed. Note that SVR+ requires careful selection of all four optimal parameters and thus, a full-grid search was

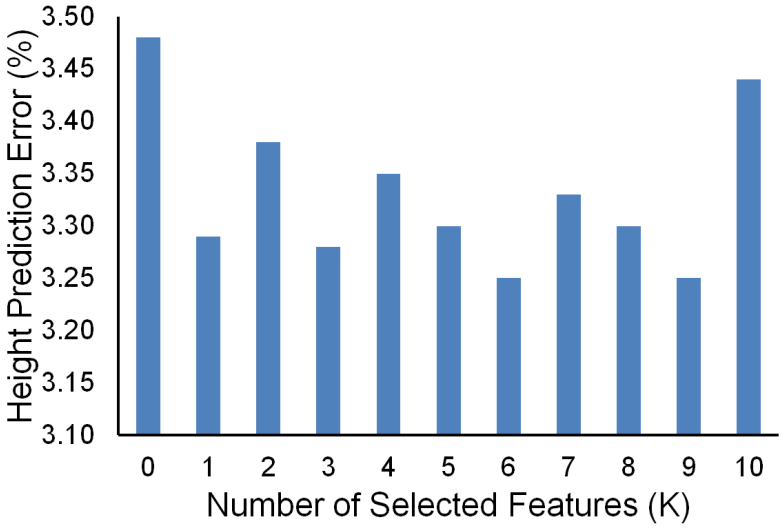

Figure 4.2: Height prediction error for different number of selected features.

performed.

Both ϵ-SVR and ϵ-SVR+ optimization problems are convex and can be solved using quadratic programming (QP). For large datasets and to enable fast training, a sequential minimal optimization (SMO) [90] technique is frequently used which divides a large QP optimization problem into a series of smaller QP problems. However, for the purposes of this work, a regular QP solver was used as provided in the CVXOPT package [2] and implemented the ϵ-SVR+ algorithm following the dual formulation described in the work of Vapnik and Vashist [116].

Regression-based Height Prediction: Three methods are evaluated: (i) a regular ϵ-SVR which predicts the height attribute using ratios of observable anthropometric measurements, (ii) the ϵ-SVR+ algorithm which leverages privileged information at training time to obtain a more accurate optimization solution in the observable space, and (iii) the proposed PIP approach which predicts the K most informative privileged features at test time and uses them with the observable space to estimate the height. Different models are trained and cross-validated per gender, and thus separate

results are reported per gender and per height quartiles (e.g., the 1^{st} quartile corresponds to the shortest 25% of the subjects). The regression error (%) results of the aforementioned techniques are depicted in Table 4.3 per gender and per quartile (Q).

Leveraging privileged information is beneficial for the estimation of soft biometric attributes such as height because both ϵ-SVR+ and PIP performed better than a regular ϵ-SVR. Moreover, when the gender of the human is not known beforehand, which would be the case in a real-life biometric application, the proposed approach outperformed the other two techniques in all but one case. Especially in the first and the fourth quartiles, which proved to be the most challenging ones, PIP demonstrated smaller estimation error than the other two techniques. The most challenging quartiles contain either the shortest samples (first female quartile) or the tallest subjects (fourth male quartile). The reason for this is that these groups contain heights that are close to the boundaries of the range of height values and are difficult to predict by a universal model. On the contrary, the height estimation of samples belonging to the second and the third quartiles had the smallest error in all cases. The overall performance between males and females appears to be approximately the same. Finally, using the gender of a human as prior information can reduce the height estimation error compared to a scenario in which the dataset comprises samples of both genders. Note that a 1% error corresponds approximately to a 1.6 *cm* absolute difference between the estimated and the actual value.

Selecting the optimal number of privileged features: From our investigation of predicting the privileged selected features at test time, two interesting questions arise: (i) what is considered privileged information, and (ii) what is the optimal number (K) of features to be selected that leads to best prediction accuracy? Although conceptually it might seem reasonable to use circumferences of human limbs as prior information to boost the height prediction accuracy, this is possible only through careful selection of the parameters. Second, K was selected by experimenting with

different values, while concurrently performing cross-validation using the same parameters for all models to reduce the training computational time, and then estimated the height for each value. Note that our goal was to find the smallest number of K, thus, from the obtained results depicted in Figure 4.2 K was set equal to six. A limitation of the mRMR algorithm [87], is that it does not consider information from groups of features but ranks them individually, which explains the fluctuations in the obtained error for a different number of selected features.

4.3 Curriculum Learning of Visual Attribute Clusters for Multi-Task Classification

4.3.1 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.

4.3.1.1 Multi-label ConvNet architecture

To mitigate the lack of training data we employ the pre-trained VGG-16 [104] network. VGG-16, is the network from Simonyan and Zisserman [104] which was one of the first methods to demonstrate

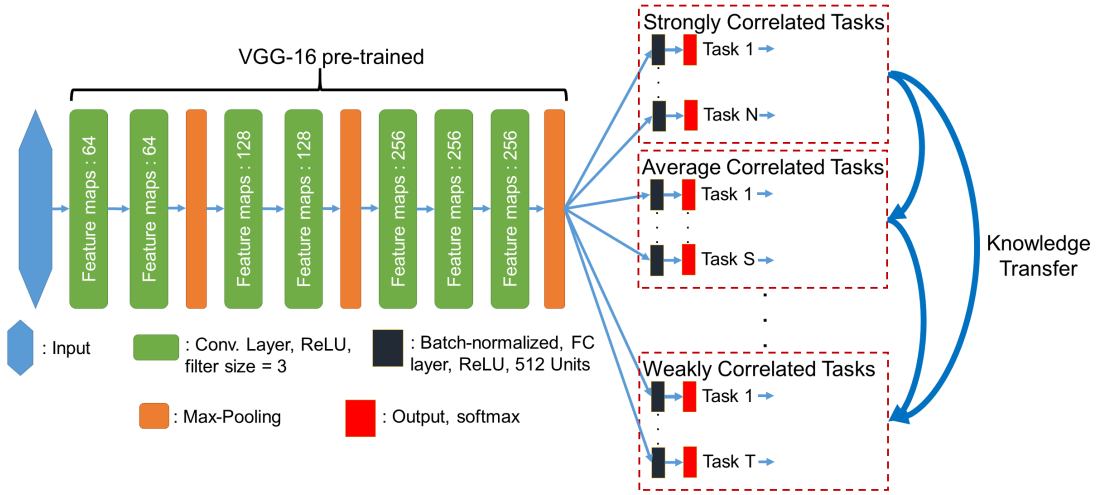

Figure 4.3: Architecture of the ConvNet used for several groups of tasks. The VGG-16 pre-trained part is kept frozen during training and only the weights of the last layers are learned. The different groups of tasks are learned sequentially using a curriculum learning paradigm. However, when the latter groups of tasks are trained, the tasks which have already been learned, contribute to the total cost function (Figure best viewed in color.)

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 [93], 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 4.3. 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 [133] and thus, not transferable. Following that, for every task we added a batch-normalized [44] 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 ting. Although shuffling and normalizing each batch has proven to reduce the need of dropout, we observed that adding a dropout layer [107] 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.* [112], 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 pixel values as proposed by Krizhevsky *et al.* [56]. 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.*[37], proved to be very effective as it reduced the variation on the predictions.

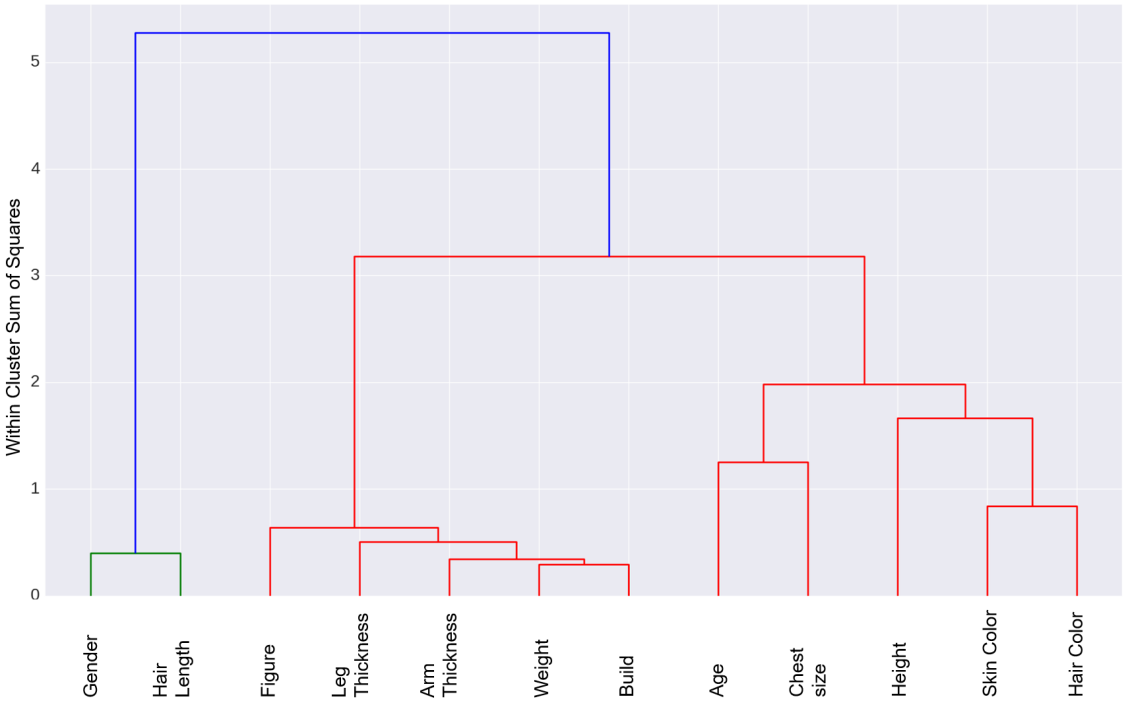

Figure 4.4: Dendrogram illustrating the arrangement of clusters.

4.3.1.2 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 $t_i, i = 1...T$ 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:

$$\underset{S(t_i)}{\text{minimize}} \frac{1}{T} \sum_{j=1}^{T} \mathcal{E}(\hat{y}_{t_j}, y_{t_j}) \quad , \tag{4.4}$$

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.

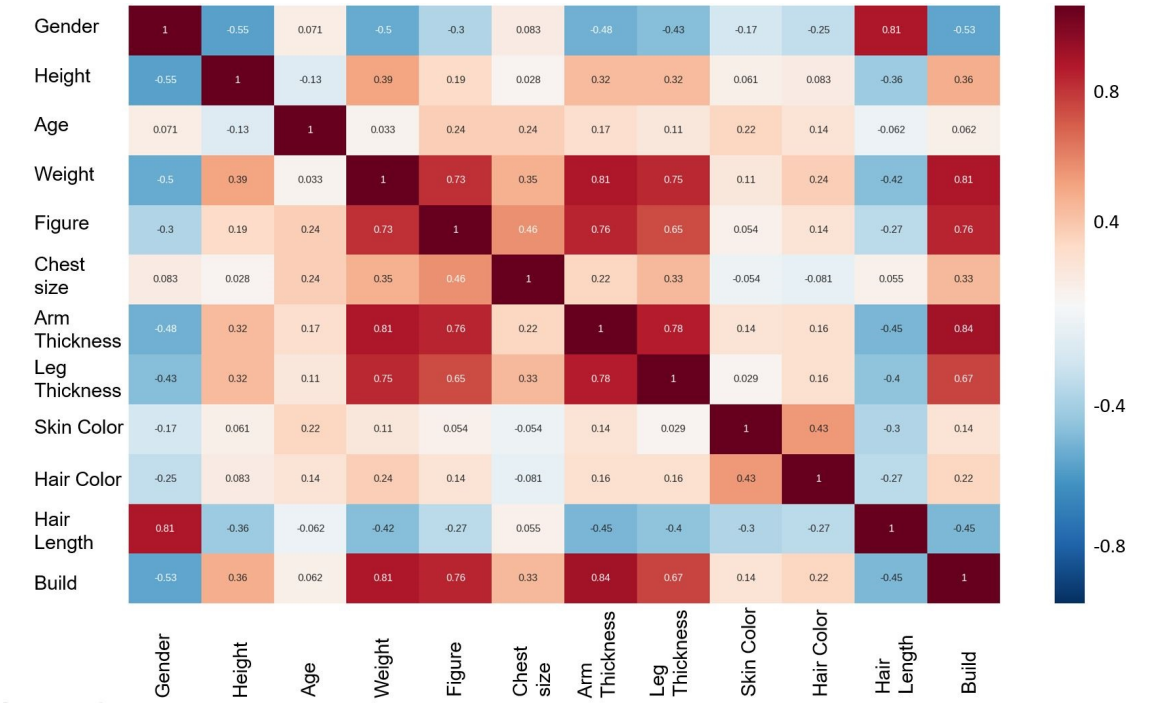

Figure 4.5: Pairwise correlation matrix between the visual attributes of the SoBiR dataset.

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 measurements in other clusters.

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. It is a variance-minimizing approach and which resembles the k-means algorithm but tackled with an agglomerative hierarchical approach. 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 existing (but yet unused) cluster v is defined as:

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 + \frac{|v|}{T}d(s,t)^2}\,, \tag{4.5}$$

where s, t are the clusters which are joined into cluster u, and $T = |v|+|s|+|t|$. Ward [125], 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 [79] in the form of a dendrogram is depicted in Figures 4.4 and 4.5. 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

Algorithm 4.2: Finding the learning sequence of attribute clusters

Input : Training labels Y, WCSS threshold τ

$P \leftarrow$ compute Pearson correlation coefficient matrix split based on labels Y

$G \leftarrow$ split into clusters using Eq. (4.5) along with P, labels Y, and τ

for *group* g_i *in* G **do**

  $S_i \leftarrow$ compute average of cross-correlation within g_i using Eq. (4.6)

end

$S(g_i) \leftarrow$ compute learning sequence of clusters by sorting S_i's in a descending order

Output: Learning sequence of clusters of visual attributes $S(g_i)$

time only within the cluster as follows:

$$p_{i,c} = \sum_{j=1, j\neq i}^{T} \frac{cov(y_{t_{i,c}}, y_{t_{j,c}})}{\sigma(y_{t_{i,c}})\sigma(y_{t_{j,c}})},\ i = 1, ..., T\ , \tag{4.6}$$

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. 4.2, 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 4.3: Multi-task curriculum learning training

Input : Training set X, training labels Y, learning sequence of clusters $S(g_i)$ from Algorithm 4.2

for *group* g_i *in* $S(g_i)$ **do**

  Initialize C_i from rest of already trained groups of tasks (if any)

  $C_i \leftarrow$ train model using (X, Y_i) by minimizing the loss in Eq. (4.8)

end

Output: Parameters of network containing all groups of tasks

4.3.1.3 Multi-Task Curriculum Learning

In the scenario we are investigating, we solve multiple binary unbalanced classification tasks simultaneously. The proposed learning paradigm is described in Algorithm 4.3.

Similar to Zhu *et al.* [143], we employ the categorical cross-entropy function between predictions and targets, which for a single attribute t is defined as follows:

$$L_t = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{1/M_j}{\sum_{n=1}^{M} 1/M_n}\right)\cdot \mathbb{1}[y_i = j]\cdot \log(p_{i,j}), \tag{4.7}$$

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.* [136] we leverage the knowledge learned by backpropagating the following loss:

$$L_j = \lambda \cdot L_t + (1 - \lambda) \cdot L_j^f, \tag{4.8}$$

where L_j^f is the total loss computed during the forward pass using Eq. (4.7) 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.

4.3.2 Experiments

4.3.2.1 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 [79], VIPeR [32], and PETA [20] datasets. The selected datasets are of varying difficulty and contain different visual attributes and training set sizes. In each dataset, we follow the same evaluation protocol with the rest of the literature.

SoBiR dataset: The recently introduced SoBiR dataset [79] contains 800 images of 100 people. The dimensions of each image are 256×256. The SoBiR dataset comprises 12 soft biometric labels (e.g., gender, weight, age, height) and four forms of comprehensive human annotation (absolute versus relative and categorical versus binary). In our experimental investigation, we used the comparative binary ground-truth annotations (e.g., taller/shorter instead 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 [79, 86]; 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.

PETA dataset: The PETA dataset [20] 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 [143], highly imbalanced attributes are discarded and the remaining 45 binary visual attributes are employed.

4.3.2.2 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 $\tau = 1.9$, 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. In addition, we performed SVM with features extracted from the last fully-connected layer of the pre-trained VGG-16 network and the obtained results are provided in the third column of Table 4.4. Feature vectors $4,096 \times 1$ were extracted for each image, and an SVM was trained using the optimal parameters obtained from the validation set. 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 4.4. 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 [129, 62]. In Figure 4.6, 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.

4.3.2.3 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. 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

Table 4.4: Classification accuracy of different learning paradigms on the SoBiR dataset. In CILICIA, four clusters were formed and attributes are in descending order based on their intra cross-correlation. Results highlighted with light blue indicate statistically significant improvement using the paired-sample t-test.

Soft Label	SVM with Handcrafted Features	SVM with Deep Features	Individual Learning	Multi-Task Learning	CILICIA
Gender	72.1	74.5	80.4	79.6	**85.2**
Height	64.7	61.8	73.9	72.0	**77.0**
Age	58.5	55.3	62.6	61.9	**64.5**
Weight	57.7	65.3	67.7	71.0	**74.1**
Figure	57.8	64.3	**68.7**	67.1	67.3
Chest size	58.7	54.5	64.9	**68.9**	67.5
Arm thickness	60.1	70.5	72.0	73.1	**73.7**
Leg thickness	56.7	65.5	68.9	71.0	**72.6**
Hair length	71.8	72.5	78.9	79.2	**85.9**
Muscle build	58.5	66.3	73.3	74.5	**75.8**
Average	61.9	63.6	71.0	71.3	**74.2**

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.* [143].

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

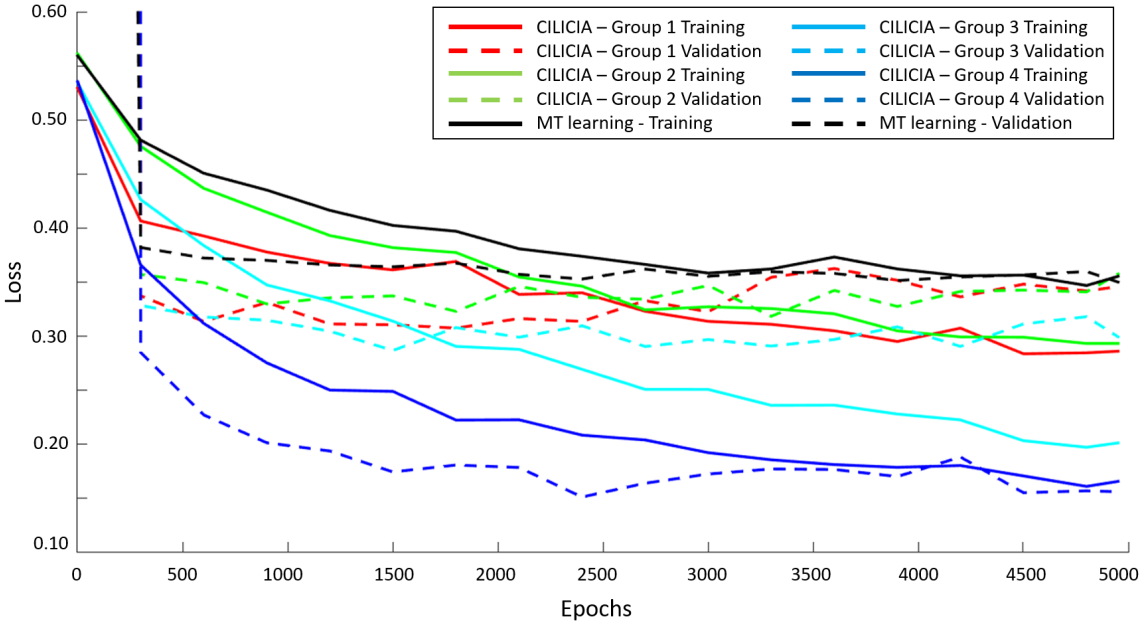

Figure 4.6: Convergence plot for all the groups of CILICIA as well as Multi-Task learning on the SoBiR dataset.

report our results in comparison with those of Zhu *et al.* [143] after grouping the attributes in Table 4.5. The imbalance ratio in this dataset, 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. 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.* [143] 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 4.7 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.

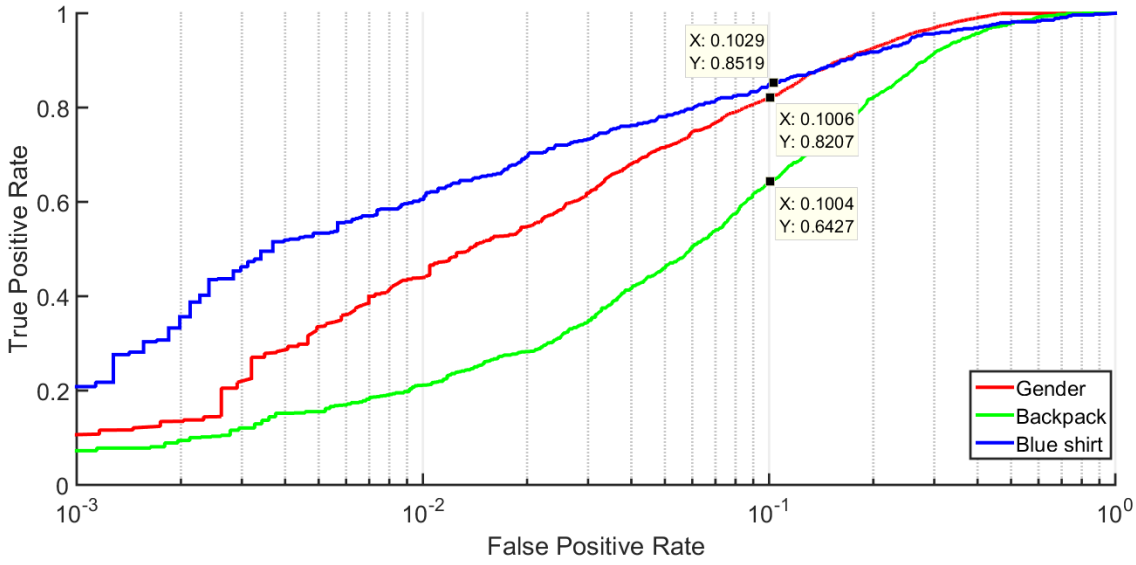

Figure 4.7: ROC curves for the visual attributes of "gender", "blue shirt", and "has backpack". The x-axis is in semi-logarithmic scale and the depicted values correspond to the recall rate (%) when the false positive rate is 10%.

4.3.3 Ablation Studies

Why is Knowledge Transfer important? To assess the impact of transferring knowledge from groups of tasks which have already converged to ones 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 classification frameworks. We report the obtained results in the last two columns of Table 4.6. 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. Transferring knowledge from a strongly correlated group

Table 4.5: Performance comparison on the PETA dataset for different types of attributes.

Visual Attribute	Imbalance Ratio	Accuracy (%)		Recall rate (%) @FPR = 10%		AUC (%)	
		Zhu *et al.* [143]	CILICIA	Zhu *et al.* [143]	CILICIA	Zhu *et al.* [143]	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**

of tasks to the weakly improves the performance of the latter by 1.89% compared to a typical multi-task classification learning framework.

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 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 4.6. 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. The obtained results are in line with previous methods that can be found in the literature [43, 36] that have exploited label correlations to improve multi-task learning.

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 4.4 and 4.6) which led 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 4.6). We then investigated whether the work of Bengio *et al.* [5], 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.* [5]. In addition, note that when Bengio *et al.* [5] introduced curriculum learning after they defined an entropy-based curriculum they demonstrated that introducing gradually more difficult examples speeds-up online training.

Table 4.6: Ablation experiments to assess the effectiveness of knowledge transfer and correlation-based split.

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**

4.3.4 Performance Analysis and Limitations

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. This was 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 that such as the GoogLeNet [113], or the Highway Networks [108]. 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 [76, 136]. 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.

4.4 Deep Imbalanced Attribute Classification using Visual Attention Aggregation

Given an image of a human our goal is to predict its visual attributes. More formally, our input consists of an image x along with its corresponding labels $y = [y^1, y^2, \ldots, y^C]^T$ where C is the

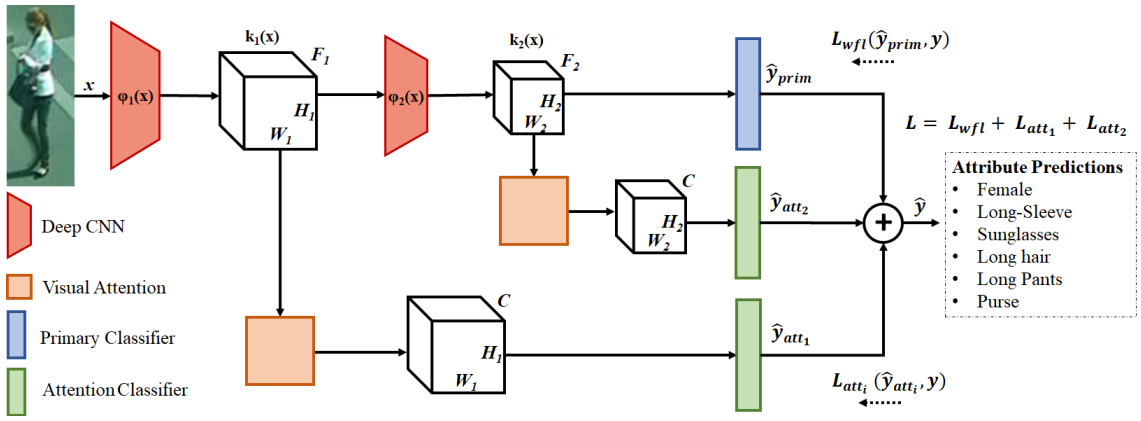

Figure 4.8: Given an image of a human we aspire to predict C visual attributes. Visual attention mechanisms are placed in two different levels of the network to identify spatial information that is relevant to each attribute with only attribute-level supervision. The predictions from the attention and the primary classifiers are aggregated at a score level and the whole network is trained end-to-end with two loss functions that handle class imbalance and hard samples ($\mathcal{L}_{wfl}$) and penalize attention masks with high-prediction variance ($\mathcal{L}_{att}$).

total number of attributes and y^c a binary label that indicates the presence or absence of a particular attribute in the image. We aspire to learn a function $f(x)$ that minimizes the prediction error over all attributes:

$$f^* = \min_f \sum_{c=1}^{C} \mathcal{L}\left(f^c(x), y^c\right), \tag{4.9}$$

where $\mathcal{L}$ corresponds to the selected loss function (e.g., binary cross-entropy loss) and f^* is a deep neural network.

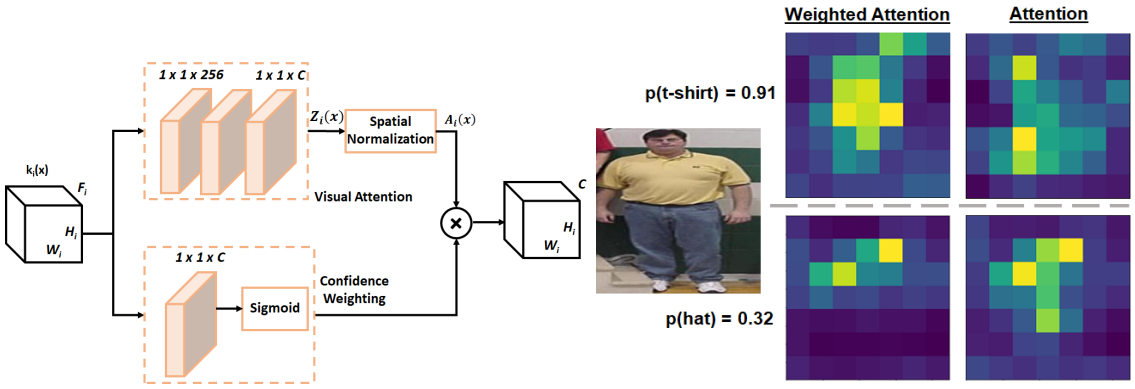

Figure 4.9: Our attention mechanism (upper-left) maps feature representations of spatial resolution $H_i \times W_i$ and F_i channels to C channels (one for each attribute) with the same size which are then spatially normalized to enforce the model to focus its resources to the most relevant region of the image. The attention masks are weighted by attribute confidences (lower-left) which as we demonstrate on the right, apply larger weights to the attribute-corresponding areas. For example, more emphasis is given in the middle-upper part when looking for a t-shirt and to the upper part of the image when looking for a hat (even when it is not there).

4.4.0.1 Multi-scale Visual Attention and Aggregation

In this work, we experimented with both ResNets [37] and DenseNets [40] as backbone architectures and thus, we opted for the representations after the third and the fourth stage/block of layers. The concept of extracting attention information can be expanded to more spatial resolutions/scales besides two at the expense of learning additional parameters. We will thus refer to the first part of the networks (up to stage/block three) as $\phi_1(\cdot)$ and to the part from there and until the classifier as $\phi_2(\cdot)$. In our primary network, which unless otherwise specified is a ResNet-101 architecture

(deep CNN module in Figure 4.8), given an image x, we obtain three-dimensional feature representations:

$$k_1(x) = \phi_1(x), \;\; k_1(x) \in \mathcal{R}^{H_1 \times W_1 \times F_1}, k_2(x) = \phi_2(k_1(x)), \;\; k_2(x) \in \mathcal{R}^{H_2 \times W_2 \times F_2}. \quad (4.10)$$

For 224×224 images the attention mechanism is placed to features of channel size F_i equal to $1,024$ and $2,048$ with spatial resolutions $H_i \times W_i$ equal to 14×14 and 7×7 respectively. Finally, the classifier of the primary network outputs logits $\hat{y}_{prim}(x) = W_{prim} k_2(x) + b_{prim}$ where (W_{prim}, b_{prim}) are the parameters of the classification layer.

With simplicity in mind, our attention mechanism, depicted in Figure 4.9, consists of three stacked convolutional layers (along with batch-normalization and ReLU) with a kernel size equal to one. Due to the multi-label nature of the problem, the last convolutional layer maps the channels to the C number of classes (i.e. attributes). This is different than most attention works (with one label per image) that extract saliency maps of the same spatial/channel size of the given feature representation. The attribute-specific attention maps $z^c_{h,w}$ are then spatially normalized to $a^c_{h,w}$using a spatial softmax operation:

$$a^c_{h,w} = \frac{\exp(z^c_{h,w})}{\sum_{h,w} \exp(z^c_{h,w})}, \quad (4.11)$$

where h, w correspond to the height and width dimension and c to the corresponding attribute label. The spatial softmax operation results in attention masks with the property $\sum_{h,w} a^c_{h,w} = 1$ for each attribute c and is used to enforce the model to focus its resources to the most relevant region of the image. We will refer to the attention mechanism comprising the three convolutional layers as $\mathcal{A}$ and thus, for each spatial resolution i we first obtain unnormalized attentions $Z_i(x) = \mathcal{A}(k_i(x))$, which are then spatially normalized using Eq. (4.11) resulting in normalized attention masks $A_i(x)$.

Following the work of Zhu *et al.* [142], we concurrently pass the feature representations to

a single convolutional layer with C channels (same as the number of classes) followed by a sigmoid function. The role of this branch is to assign weights to the attention maps based on label confidences and avoid learning from the attention masks when the label is absent. The weighted attention maps reflect both attribute information at different spatial locations and label confidences. We observed in our experiments that this confidence-weighting branch boosts the performance by a small amount and helps the attention mechanism learn better saliency heatmaps (Figure 4.9 right).

Combining the output saliency masks from different scales can be done either at a prediction level (i.e. averaging the logits) or at a feature level [123]. However, aggregating the attention masks at a feature level provided consistently inferior performance. We believe that this is because the two attention mechanisms extract masks that give emphasis to different spatial regions that when added together fail to provide the classifier with attribute-discriminative information. Thus, we opted for the former approach and fed each confidence-weighted attention mask to a classifier to obtain logits $\hat{y}_{att_i}$ of the attention module i. The final attribute predictions of dimensionality $1 \times C$ for an image x are then defined as $\hat{y} = {}^{(\hat{y}_{prim}+\hat{y}_{att_1}+\hat{y}_{att_2})}/_{3}$.

4.4.1 Deep Imbalanced Classification

Using the output predictions of the primary model $\hat{y}_{prim}$ which have the same dimensionality $1 \times C$ (i.e. one for each attribute), a straight-forward approach adopted by Zhu *et al.* [142] is to train the whole network using the binary cross-entropy loss $\mathcal{L}_{bce}$ as:

$$\mathcal{L}_{bce}(\hat{y}_{prim}, y) = -\sum_{c=1}^{C} \log(\sigma(\hat{y}^c_{prim}))y^c + \log(1 - \sigma(\hat{y}^c_{prim}))(1 - y^c), \tag{4.12}$$

where $(\hat{y}^c_{prim}, y^c)$ correspond to the logit and ground-truth labels for attribute c, and $\sigma(\cdot)$ is the sigmoid activation function. However, such a loss function ignores completely the class imbalance. Aiming to alleviate this problem both at a class- and at an instance-level, we propose to use for our

primary model a weighted-variant of the focal loss [70] defined as:

$$\mathcal{L}_{wfl}(\hat{y}_{prim}, y) = -\sum_{c=1}^{C} w_c \Big(\big(1 - \sigma(\hat{y}^c_{prim})\big)^{\gamma} \log\big(\sigma(\hat{y}^c_{prim})\big)\, y^c \; + \\ + \sigma(\hat{y}^c_{prim})^{\gamma} \log(1 - \sigma(\hat{y}^c_{prim}))(1 - y^c) \Big), \quad (4.13)$$

where γ a hyper-parameter (set to 0.5), which controls the instance-level weighting based on the current prediction giving emphasis to the hard misclassified samples, and $w_c = e^{-a_c}$, where a_c the prior class distribution of the c^{th} attribute as in [99].

Unlike the face attention networks [121], which learn the attention masks based on ground-truth facial bounding boxes, in the human attribute domain such information is not available. This means that the attention masks will be learned based on attribute-level supervisions y. The attention masks of dimensionality $H_i \times W_i \times F_i$ are fed to a classifier which outputs logits $\hat{y}_{att_i}$ for each spatial resolution i. To account for the weak supervision of the attention network, we decided to focus on the attention masks with high prediction variance. Similar to the work of Chang *et al.* [9], after some burn-in epochs in which $\mathcal{L}_{bce}$ is used, we start collecting the history H of the predictions $p_H(y_s|x_s)$ for the s^{th} sample and compute the standard deviation across time for each sample within the batch:

$$\widehat{std}_s(H) = \sqrt{\widehat{var}\big(p_{H^{t-1}}(y_s|x_s)\big) + \frac{\widehat{var}\big(p_{H^{t-1}}(y_s|x_s)\big)^2}{|H_s^{t-1}| - 1}}, \quad (4.14)$$

where t corresponds to the current epoch, $\widehat{var}$ to the prediction variance estimated in history H^{t-1} and $|H_s^{t-1}|$ the number of stored prediction probabilities. The loss for the attention-masks at level i with attribute-level supervision for each sample s is defined as:

$$\mathcal{L}_{att_i}(\hat{y}_{att_i}, y) = \big(1 + \widehat{std}_s(H)\big)\mathcal{L}_{bce}(\hat{y}_{att_i}, y). \quad (4.15)$$

Attention mask predictions with high standard deviation across time will be given higher weights in order to guide the network to learn those uncertain samples. Note that for memory reasons, our history comprises only the last five epochs and not the entire history of predictions. Finally, the total loss that is used to train our network end-to-end (the primary network and the two attention modules) is defined as:

$$\mathcal{L} = \mathcal{L}_{wfl} + \mathcal{L}_{att_1} + \mathcal{L}_{att_2}, \tag{4.16}$$

where $\mathcal{L}_{att_1}$ is applied to the first attention module that extracts saliency maps of spatial resolution 14×14, and $\mathcal{L}_{att_2}$ is similarly applied to the second attention module after the fourth stage of the primary network with spatial resolution of 7×7. Disentangling the two loss functions enables us to separately focus on different types of challenges. The weighted focal loss $\mathcal{L}_{wfl}$, handles the prior class imbalance per attribute using the weight w_c and at the same time focuses on hard misclassified positive samples via the instance-level weights of the focal loss. The attention loss $\mathcal{L}_{att}$, penalizes predictions that originate from attention masks with high prediction variance.

4.4.2 Experiments

To assess our method we performed experiments and ablation studies on the publicly available WIDER-Attribute [67] and PETA [20] datasets, which are the most widely used in this domain. The training details for both datasets are provided in the supplementary material.

4.4.2.1 Results on WIDER-Attribute

Dataset Description and Evaluation Metrics: The WIDER-Attribute [67] dataset contains 13,789 images with 57,524 bounding boxes of humans with 14 binary attribute annotations each. Besides "gender", which is balanced, the rest of the attributes demonstrate class imbalance, which can reach

Table 4.7: Evaluation of the proposed approach against nine state-of-the-art methods.

Method	Male	Long hair	Sunglasses	Hat	T-shirt	Long sleeve	Formal	Shorts	Jeans	Long Pants	Skirt	Face Mask	Logo	Plaid	**mAP**
Imbalance Ratio	1:1	1:3	1:18	1:3	1:4	1:1	1:13	1:6	1:11	1:2	1:9	1:28	1:3	1:18	
RCNN [28]	94	81	60	91	76	94	78	89	68	96	80	72	87	55	80.0
R*CNN [30]	94	82	62	91	76	95	79	89	68	96	80	73	87	56	80.5
DHC [67]	94	82	64	92	78	95	80	90	69	96	81	76	88	55	81.3
VeSPA [99]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.4
CAM [34]	95	85	71	**94**	78	96	81	89	75	96	81	73	88	60	82.9
ResNet-101 [37]	94	85	69	91	80	96	83	91	78	95	82	74	89	65	83.7
ResNet-101+MTL	94	86	68	91	81	**96**	83	91	79	95	83	74	**90**	65	83.8
ResNet-101+MTL+CRL [21]	94	86	71	91	81	**96**	83	92	79	96	84	76	**90**	66	84.7
SRN [142]*	95	87	72	92	82	95	84	92	80	**96**	84	76	**90**	66	85.1
Ours	**96**	**88**	**74**	93	**83**	**96**	**85**	**93**	**81**	**96**	**85**	**78**	**90**	**68**	**86.4**

$1:18$ and $1:28$ for attributes such as "face-mask" and "sunglasses". Following the training protocol of [99, 142], we used the human bounding box as an input to our model and mean average precision (mAP) results are reported.

Baselines: We evaluate our approach against all the methods that have been tested on the WIDER-Attribute dataset, namely R-CNN [28], R*CNN [30], DHC [67], CAM [34], VeSPA [99], SRN [142], and a fine-tuned ResNet-101 network [37]. In addition, we transform the last part of the network to perform multi-task classification (MTL) by adding a fully-connected layer with 64 units for each attribute. This enables us to additionally evaluate against CRL [21] by forming triplets within the batch using class-level hard samples. Note that DHC and R*CNN leverage additional contextual information (e.g., scene context or image parts) that intuitively should boost the performance and

VeSPA, which jointly predicts the viewpoint along with the attributes, did not train its viewpoint prediction sub-network on the WIDER-Attribute dataset. In SRN [142], the validation set was included in the training (which results in 20% more training data) and samples from the test set were used to obtain an idea about the training performance. In order to allow for a fair comparison with the rest of the methods, we re-implemented their method (which is why there is an asterisk next to their work in Table 4.7) and trained it only on the training set of the WIDER-Attribute [67] dataset. The difference between the reported results and our re-implementation is 1.2 in terms of mAP which is reasonable given the access to approximately 20% less training data.

Evaluation Results: Our proposed approach achieves state-of-the-art results on the WIDER dataset by improving upon the second best work by 1.3 in terms of mAP and by 2.7 over ResNet-101 [37] which was our primary network. Our larger improvements are in imbalanced attributes such as "Sunglasses" or "Plaid" that have visual cues in the image which demonstrates the importance of handling class imbalance and using visual attention to identify important visual information in the image. DHC and R*CNN that use additional context information performed significantly worse but this is partially because they utilize smaller primary networks. Overall the proposed approach performs better than or equal than the rest of the literature in all but one attributes and comes second behind CAM [34] at recognizing hats.

4.4.2.2 Ablation Studies on WIDER

In our first ablation study (Table 4.8 - left), we investigate to what extent the primary network affects the final performance. This is because it is commonplace that as architectures become deeper, the impact of individual add-on modules becomes less significant. On the left, we report mAP results just for the primary network (w/o adding any attention mechanisms) using different backbone architectures. On the right, we investigate the additions in terms of performance for

Table 4.8: Ablation studies on the WIDER dataset to assess the impact of individual modules on the final performance of our method.

Primary Net	**Params**	**mAP**
ResNet-50	25.6×10^6	82.3
DenseNet-121	8.1×10^6	82.9
ResNet-101	44.7×10^6	83.7
ResNet-152	60.4×10^6	84.2
DenseNet-201	20.2×10^6	84.5

Primary Net	$\mathcal{L}_{wfl}$	Attention	$\mathcal{L}_{att}$	Multi-scale	**mAP**
ResNet-101					83.7
ResNet-101	✓				84.4
ResNet-101	✓	✓			85.0
ResNet-101	✓	✓	✓		85.7
ResNet-101	✓	✓		✓	85.9
ResNet-101	✓	✓	✓	✓	86.4

attention at a single- and multi-scale level as well as the two loss functions we introduced. We observe that (i) the difference between a ResNet-50 and a DesneNet-201 architecture is more than 2% in terms of mAP, (ii) DenseNet-201, which is the highest performing primary network, is almost as good as SRN [142] due to its effective feature aggregation and reuse, and (iii) the mAP of the proposed approach is 2.1 more than the best performing primary network. In our second ablation study (Table 4.8 - right), we assess how each proposed component of our approach contributes to the final mAP. Handling class imbalance using the weighted focal loss and adding our attention mechanism just at a single scale result in mAP equal to 85.0 which performs almost as well as the existing state-of-the-art. Adding the attention loss that penalizes attention masks with high prediction variance and expanding the attention module to two scales improves the final mAP to 86.4.

Qualitative Results: In Figure 4.10, attention masks for six successful (left) and three failure cases (right) are provided. For imbalanced attributes such as sunglasses that have discriminant visual cues, the attention mechanism locates successfully the corresponding regions, which explains the

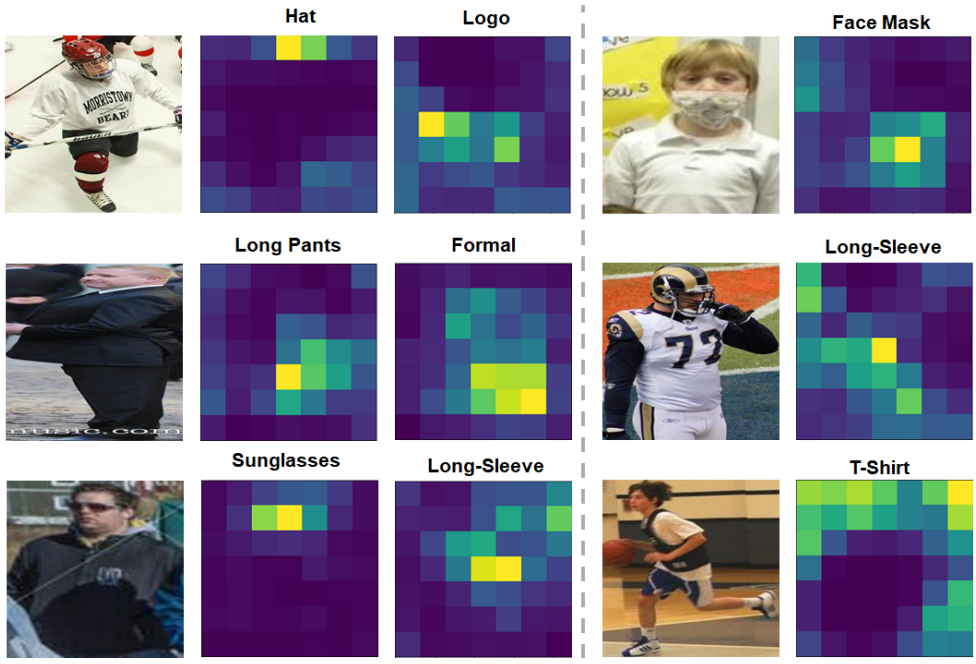

Figure 4.10: Successful attention masks (left) and failure cases (right) for attributes of the WIDER dataset. The attention masks learned are capable of finding formal clothes and long pants in the bottom part of the image, logos in the middle and sunglasses or hats in the top.

7% relative improved mAP for this attribute compared to the primary ResNet architecture.

4.4.2.3 Results on PETA

Dataset Description and Evaluation Metrics: The PETA [20] dataset is a collection of 10 person surveillance datasets and consists of 19,000 cropped images along with 61 binary and 5 multi-value attributes. We used the same train/validation/test splits with the method of Sarfraz *et al.* [99] and followed the established protocol of this dataset by reporting results on the 35 attributes for which the ratio of positive labels is higher than 5%. For the PETA dataset, two different types of metrics

Table 4.9: Evaluation of the proposed approach against 9 state-of-the-art approaches on the PETA dataset ranked by F1-score.

Method	**mA**	**Acc**	**Prec**	**Rec**	**F1**
ACN [111]	81.15	73.66	84.06	81.26	82.64
SRN [142]* (w/ $\mathcal{L}_{bce}$)	80.55	74.24	84.04	82.48	83.25
WPAL-FSPP [135]	84.16	74.62	82.66	85.16	83.40
DeepMAR [61]	82.89	75.07	83.68	83.14	83.41
GoogleNet [113]	81.98	76.06	84.78	83.97	84.37
ResNet-101 [37]	82.67	76.63	85.13	84.46	84.79
WPAL-GMP [135]	**85.50**	76.98	84.07	85.78	84.90
SRN [142]* (w/ $\mathcal{L}_{wfl}$)	82.36	75.69	85.25	84.59	84.92
VeSPA [99]	83.45	77.73	86.18	84.81	85.49
Ours	84.59	**78.56**	**86.79**	**86.12**	**86.46**

are reported namely label-based and example-based. For the label-based metrics due to the unbalanced distribution of the attributes, we used the balanced mean accuracy (mA) for each attribute that computes separately the classification accuracy of the positive and the negative examples and then computes the average. For the label-based metrics, we report accuracy, precision, recall, and F1-score averaged across all examples in the test set.

Baselines: We compared our approach with all the methods that have been tested on the PETA dataset, namely the ACN [111], DeepMAR [61], two variations of WPAL [135], VeSPA [99], the GoogleNet [113] baseline reported by Sarfraz *et al.* [99], ResNet-101 [37] and SRN [142].

Evaluation Results: From the complete evaluation results in Table 4.9, we observe that the proposed approach achieves state-of-the-art results in all example-based metrics and comes second to WPAL [135] in terms of balanced mean accuracy (mA). We believe this is due to the fact that different methods use different metrics, based on which they optimize their models. For example, our approach is optimized based on the F1 score which balances between precision and recall and is applicable in search applications. Our approach improves upon a fine-tuned ResNet-101 architecture by approximately 2% in terms of F1 score which demonstrates the importance of the visual attention mechanisms. Notably, we improve upon VeSPA [99] in all evaluation metrics despite the fact that they utilize additional viewpoint information to train their model. Finally, we observe that by using the weighted variant of focal loss ($\mathcal{L}_{wfl}$) instead of the binary-cross entropy loss ($\mathcal{L}_{bce}$), the F1 score of SRN [142] increases by 1.7%. This demonstrates why failing to account for class imbalance affects the performance of deep attribute classification models.

4.4.2.4 Ablation Studies on PETA

Based on the experimental analysis an important question arises: can similar results be obtained with significantly fewer parameters? Aiming to find out the impact of large backbone architectures in the final performance, we investigated how each component of our work performs using a pre-trained DenseNet-121 [40] architecture. DenseNet-121 contains $7.5\times$ less parameters compared to ResNet-101 due to efficient feature propagation and reuse. To our surprise, when all components are included (last row in Table 4.10), the performance drop in terms of F1 score is less than 2%. In addition a variety of feature aggregations were explored by either up-sampling the smaller attention masks, max-pooling the larger or mapping the larger to the smaller using a convolutional layer with stride equal to two. Although the latter approach performed better than up-sampling/down-sampling, the aggregation of the attention information at a logit level is superior compared to

feature level aggregation. This is because the two attention mechanisms extract masks that give emphasis to different spatial regions that when added together fail to provide the classifier with attribute-discriminative information.

Table 4.10: Ablation studies to assess the impact of each submodule to the final result using a light-weight backbone architecture with $7.5\times$ less parameters than a ResNet-101.

Primary Net	Class Weight	$\mathcal{L}_{wfl}$	Attention	Multi-scale (feature aggr.)	Multi-scale (score aggr.)	**F1**
DenseNet-121	✓					82.1
DenseNet-121	✓	✓				82.9
DenseNet-121	✓	✓	✓			83.8
DenseNet-121	✓	✓	✓	✓		84.1
DenseNet-121	✓	✓	✓		✓	84.7

4.4.2.5 Sources of Error and Further Improvements

Where does the proposed method fail and what are the characteristics of the failure cases? A significant limitation of most pedestrian attribute classification methods (including ours) is that they resize the input data to a fixed square-size resolution (e.g., 224×224) in order to feed them to deep pre-trained architectures. Human crops are usually rectangular captured from different viewpoints and thus, when resized to a square, important spatial information is lost. One possible solution to this would be feeding the whole image at a fixed resolution that does not interfere with the spatial relations and then extract human-related features using ROI-pooling at a stage within the network. To cope with the high viewpoint variance, the spatial transformer networks of Jaderberg *et al.* [46]

could be employed to align the input image before feeding it to the network, a practice which is very common in face recognition applications [88, 114, 49]. A second source of error is the very low resolution of several images especially in the PETA dataset, which makes it hard even for the human eye to identify the attribute traits of the depicted human. In addition, the provided annotations contain a third unspecified/uncertain class, which is used as negative during training in the literature, that further dilutes the learning process. Applying modern super-resolution techniques [53, 18] could alleviate this issue but only to some extent. Regarding errors due to modeling richer feature representations could be extracted using feature pyramid networks [69] since they extract high-level semantic feature maps at multiple scales. Modern visual attention mechanisms [66, 119] could be adapted to a multi-label setup and applied to achieve superior performance at the expense of a larger parameter space.

4.4.2.6 Discussion on Class Imbalance

Class imbalance is an important problem in computer vision that is overlooked by the research community. Visual attributes are largely imbalanced in nature and both datasets used in this work contained attributes that demonstrated imbalance up to 1:28. Traditional solutions include over-sampling the minority classes or under-sampling the majority classes to compensate for the imbalanced class ratio as well as cost-sensitive learning where classification errors are penalized differently. Such approaches have been extensively used in the past, but they suffer from some limitations. For example, over-sampling introduces redundant information making the models prone to overfitting [21], whereas under-sampling may remove valuable discriminative information [21]. Recent works with deep convolutional neural networks introduced a sampling procedure of triplets, quintuplets or clusters of samples that satisfy some properties in the feature-space and used them to regularize their models. However, sampling triplets is a computationally expensive

procedure and the characteristics of the triplets in a batch-mode setup might vary significantly. This work demonstrated that while assigning prior class weights can alleviate part of this problem, a weighted-variant of the focal loss works consistently better by handling imbalanced classes and at the same time focusing on hard examples.

Chapter 5

Objective 3: Text-to-image matching for Person Search

In this chapter, a novel text-to-image matching approach named TIDAM is introduced which employs an adversarial learning framework to learn better feature representations.

5.1 Methodology

5.1.1 Joint Feature Learning

During training our objective is to learn discriminative visual and textual feature representations capable of accurately retrieving the ID (or the category) of the input from another domain. The training procedure is depicted in Figure 5.1 (here a text-based person search application is used as an example) and is described in detail below. Specifically, our input at training-time consists of

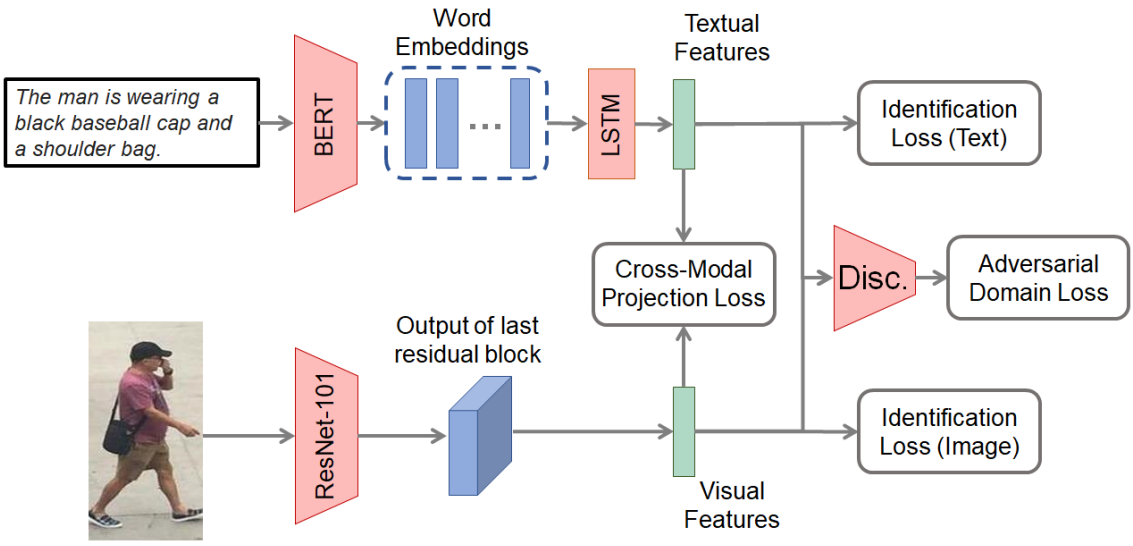

Figure 5.1: The proposed approach consists of three modules: (i) the feature extraction module which extracts textual and visual features using their corresponding backbone architectures, (ii) the identification and cross-modal projection losses that match the distributions originating from the same identity, and (iii) a domain discriminator that pushes the model to learn domain-invariant representations for effective text-image matching.

triplets (V_i, T_i, Y_i) where V_i is the image input from the visual domain V, T_i a textual description from the textual domain T describing that image, and Y_i is the identity/category of the input. To learn the visual representations denoted by $\phi(V_i)$ a ResNet-101 network is used as a backbone network. The feature map of the last residual block is projected to the dimensionality of the feature vector using global average pooling and a fully-connected layer. We opted for the original backbone architecture without any attention blocks [11, 98] in order to keep the backbones simple and easy-to-reproduce in any framework and avoid having to learn more parameters.

Learning discriminative representations from both modalities is of paramount importance for text-to-image matching. While for the image domain, most existing methods [11, 63, 140] rely on deep architectures that have demonstrated their capability of extracting discriminative features

for a wide range of tasks, this is not the case for the text domain. Prior work usually relies on a single LSTM [38] to model the textual input and learn the features that correspond to the input sentence. We argue that one of the main reasons that prevent existing computer vision methods from performing well on text-to-image matching problems is due to the fact that the textual features are not discriminative enough. To address this limitation, we borrow from the NLP community a recently proposed language representation model named BERT. The sequence of word embeddings extracted from BERT are then fed to a bidirectional LSTM [38] which effectively summarizes the content of the input textual description. Finally, the textual representation denoted by $\tau(T_i)$ is obtained by projecting the output of the LSTM to the dimensionality of the feature vector using a fully-connected layer. The reason an LSTM is employed on the output word embeddings is because it gives us the flexibility to initially freeze the weights of the language model and fine-tune only the LSTM along with the fully-connected layer and thus, significantly reducing the number of parameters. Once an adequate performance is observed, we unfreeze the weights of the language model and the whole network is trained end-to-end.

5.1.2 Cross-Modal Matching

Given the visual and textual features, our aim is to introduce loss functions that will bring the features originating from the same identity/category close together and push away features originating from different identities. To accomplish this task we introduce two loss functions for identification and cross-modal matching. The identification loss is a norm-softmax cross entropy loss [72, 120] that introduces an L_2-normalization on the weights of the output layer. By doing so, it enforces the model to focus on the angle between the weights of the different samples to perform identification instead of their magnitude. For the visual features, the norm-softmax cross entropy loss can be

described as follows:

$$L_I^v = \frac{1}{B}\sum_{i=1}^{B} -\log\left(\frac{\exp(W_i^T\phi(V_i)+b_i)}{\sum_j \exp(W_j^T\phi(V_i)+b_j)}\right), s.t. ||W_j|| = 1 \ , \tag{5.1}$$

where B corresponds to the batch size and W_i, b_i are the weights and the bias of the classification layer for the visual feature representation $\phi(V_i)$. The loss for the textual features L_I^t is computed in a similar manner and the final classification loss for identification $L_I = L_I^t + L_I^v$. Note that for datasets that do not have ID labels but only image-text pairs (e.g., the Flickr30K dataset [91]), we assign a unique ID to each image and use that ID as ground-truth for the identification loss. However, focusing solely at performing accurate identification is not sufficient for cross-modal matching since no connection between the representations of the two modalities has been introduced thus far. Towards this direction, we use the cross-modal projection matching loss [138] which incorporates the cross-modal projection into KL divergence to associate the representations across different modalities. The text representation is first normalized $\bar{\tau}(T_j) = \frac{\tau(T_j)}{||\tau(T_j)||}$ and then the probability of matching $\phi(V_i)$ to $\bar{\tau}(T_j)$ is given by:

$$p_{i,j} = \frac{\exp\left(\phi(V_i)^T\bar{\tau}(T_j)\right)}{\sum_{k=1}^{B}\exp\left(\phi(V_i)^T\bar{\tau}(T_k)\right)} \ . \tag{5.2}$$

The multiplication between the transposed image embedding and the normalized textual embedding reflects the scalar projection between $\phi(V_i)$ onto $\bar{\tau}(T_j)$, while the probability $p_{i,j}$ represents the proportion of this scalar projection among all scalar projections between pairs in a batch. Thus, the more similar the image embedding is to the textual embedding, the larger the scalar projection is from the former to the latter. Since in each mini-batch there might be more than one positive matches (i.e. visual and textual features originating from the same identity) the true matching probability is normalized as follows:

$$q_{i,j} = \frac{Y_{i,j}}{\sum_{k=1}^{B} Y_{i,k}} \ . \tag{5.3}$$

The cross-modal projection matching loss of associating $\phi(V_i)$ with correctly matched text features is then defined as the KL divergence from the true matching distribution q_i to the probability of matching p_i. For each batch (B) this loss is defined as:

$$L_M^v = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right), \tag{5.4}$$

where ϵ is a very small number. The same procedure is followed to perform the opposite matching (i.e. from text to image to compute loss L_M^t) and the summation of the two individual losses constitutes the cross-modal projection-matching loss $L_M = L_M^v + L_M^t$.

5.1.3 Adversarial Domain Learning

When training domain adversarial neural networks [8, 27, 115] a two-player minimax game is played between a domain discriminator D and a feature generator G. Both G and D are jointly trained so as G tries to fool D and D tries to make accurate domain predictions. For the text-to-image matching problem, the two backbone architectures discussed in Section 5.1.1 serve as the feature generators G^v and G^t for the visual and textual domains that produce feature representations $\phi(V_i)$ and $\tau(T_i)$, respectively. The key idea is to learn a good general representation for each input domain that maximizes the matching performance, yet obscure the domain information. By learning to fool the domain discriminator, better feature representations are learned capable of performing text-to-image matching. The generated embeddings are fed to the domain discriminator, which classifies whether the input feature representation is drawn from the visual or the textual domain. The domain discriminator consists of two fully-connected layers that reduce the embedding size to a scalar value which is used to predict the input domain. The domain discriminator is optimized according to the following GAN [31] loss function:

$$L_D = -\mathop{\mathbb{E}}_{V_i\sim V}[\log D\left(\phi(V_i)\right)] - \mathop{\mathbb{E}}_{T_i\sim T}[\log\left(1 - D\left(\tau(T_i)\right)\right)], \tag{5.5}$$

where V and T correspond to the image and text domains respectively where samples are drawn and fed through the backbone architectures.

5.1.4 Training and Testing Details

The loss function that is used to train our network is the summation of two identification losses (L_I), the two cross-modal matching losses (L_M) and the adversarial loss of the domain discriminator (L_D):

$$L = L_I + L_M + L_D \ . \tag{5.6}$$

We used stochastic gradient descent (SGD) with momentum equal to 0.9 to train the image and discriminator networks and the Adam optimizer [54] for the textual networks. The learning rate was set to 2×10^{-4} and was divided by ten when the loss plateaued at the validation set until 2×10^{-6}. The batch-size was set to 64 and the weight decay to 4×10^{-4}. The hidden dimension of the bidirectional LSTM was equal to 512 and the dimensionality of all feature vectors was set to 512. Finally, to properly balance the training between G^v, G^t, and D we followed several of the tricks discussed by Chintala *et al.* [15].

At testing time given a textual description as a probe, its textual features ($\tau(T_i)$ extracted through the language backbone) and their distance between all image features ($\phi(V_j)$ extracted from the image backbone) in the test set is computed using the cosine similarity:

$$\cos(\theta) = \frac{\tau(T_i)\, \phi(V_j)}{||\tau(T_i)||\, ||\phi(V_j)||}. \tag{5.7}$$

The distances are then sorted and rank-1 through rank-10 results are reported. For image-to-text matching the same process is followed by using the image features as probe and retrieving the most relevant textual descriptions.

5.2 Experiments

Datasets: To evaluate our method, four widely-used publicly available datasets were used and their evaluation protocols were strictly followed. We opted for these datasets in order to test TIDAM on a wide range of tasks ranging from pedestrians and flowers to objects and scenes. TIDAM was tested on (i) the CUHK-PEDES [64] that contains images of pedestrians accompanied by two textual descriptions, (ii) the Flickr30K dataset [91] which contains a wide variety of images (humans, animals, objects, scenes) with five descriptions for each image, (iii) the Caltech-UCSD Birds (CUB) [92] dataset that consists of images of birds with 10 descriptions for each image and finally, (iv) the Flowers [92] dataset that consists of images of flowers originating for 102 categories with 10 descriptions for each image.

Evaluation Metrics: The evaluation metrics used in each dataset are adopted. Thus, for the CUHK-PEDES and Flickr30K datasets rank-1, rank-5, and rank-10 results are presented for each method. For the CUB and Flowers datasets, the $AP@50$ metric is utilized for text-to-image retrieval and rank-1 for image-to-text matching. Given a query textual class, the algorithm first computes the percentage of top-50 retrieved images whose identity matches that of the textual query class. The average matching percentage of all test classes is denoted as $AP@50$. Finally, note that in each dataset we compare TIDAM with the four to seven best-performing methods.

5.2.1 Quantitative Results

CUHK-PEDES Dataset: We evaluate our approach against the seven best-performing methods that have been tested on the CUHK-PEDES dataset and present text-to-image matching results in Table 5.1. Some key methods that have been evaluated on this dataset include (i) IATV [63] which learns discriminative features using two attention modules working on the both modalities

Table 5.1: Text-to-image results (%) on the CUHK-PEDES dataset. Results are ranked based on the rank-1 accuracy.

Method	Rank-1	Rank-5	Rank-10
GNA-RNN [64]	19.05	-	53.64
IATV [63]	25.94	-	60.48
PWM-ATH [13]	27.14	49.45	61.02
GLA [11]	43.58	66.93	76.26
Dual Path [140]	44.40	66.26	75.07
CMPM + CMPC [138]	49.37	-	79.27
TIDAM	**54.51**	**77.56**	**84.78**

at different levels but it is not end-to-end; (ii) GLA [11] which identifies local textual phrases and aims to find the corresponding image regions using an attention mechanism; (iii) and CMPM [138] in which two projection losses are proposed to learn features for text-to-image matching. TIDAM outperforms all previous works by a large margin. We observe an absolute improvement of more than 5% in terms of rank-1 over the previous best performing method [138] which originates from learning better feature representations through the identification and cross-modal matching losses as well as the proposed adversarial domain learning framework.

CUB and Flowers Datasets: We test TIDAM against the top-4 best-performing methods evaluated on these datasets and present our matching results in Table 5.2. Our method achieves state-of-the-art results in both image-to-text and text-to-image matching in both datasets. We observe performance increases of 2.2% and 3.4% in terms of rank-1 accuracy as well as 3.6% and 2.4% in terms of AP@50.

Flickr30K Dataset: We report cross-modal retrieval results on the Flickr30K dataset against the

Table 5.2: Cross-modal matching results on the CUB and Flowers datasets. The results are ranked based on the text-to-image AP@50 performance.

Method	**CUB**		**Flowers**	
	Img2Txt	Txt2Img	Img2Txt	Txt2Img
	Rank-1	AP@50	Rank-1	AP@50
Word CNN-RNN [92]	56.8	48.7	65.6	59.6
Triplet Loss [63]	52.5	52.4	64.3	64.9
IATV [63]	61.5	57.6	68.9	69.7
CMPM+CMPC [138]	64.3	67.9	68.4	70.1
TIDAM	**67.7**	**70.3**	**70.6**	**73.7**

top-6 best-performing methods. Similar to the three best-performing methods, and only in this dataset, a ResNet-152 is employed, which is why we also report the image backbone used, to allow for a fairer comparison. TIDAM surpasses all methods by a large margin in text-to-image matching and comes second only to DAN [84] in image-to-text matching. DAN employs multi-step attention blocks and thus, is able to learn "where to look" in an image which results into better image features. Unlike the rest of the datasets that contain a single primary object (i.e. flowers/birds/pedestrians only), Flickr30K contains a wide range of primary components ranging from humans and animals to objects and scenes. This image variance coupled with the relatively small number of training images make cross-modal matching a challenging task. While our approach increases the rank-1 text-to image matching by 3.2% and is capable of learning correct associations between images and descriptions as demonstrated in the qualitative results presented in Figure 5.2, there is still room for further improvements by future research.

Table 5.3: Ablation studies on the CUHK-PEDES dataset to assess the impact of individual modules on the final performance of our method.

$\mathcal{L}_I$	$\mathcal{L}_C$	BERT	ARL	Rank-1	Rank-10
✓				40.1	70.1
	✓			44.9	77.7
✓	✓			49.8	81.5
✓	✓		✓	51.3	82.4
✓	✓	✓		52.9	83.5
✓	✓	✓	✓	54.5	84.8

5.2.2 Ablation Studies

Impact of Proposed Components: In our first ablation study (Table 5.3), we assess how each proposed component of our approach contributes to the final text-to-image matching performance on the CUHK-PEDES dataset. We investigate the additions in terms of rank-1 and rank-10 accuracy for the identification ($\mathcal{L}_I$) and cross-modal projection ($\mathcal{L}_M$) losses, the addition of BERT as a backbone architecture for language modeling, and the adversarial representation learning paradigm. We observe that the identification ($\mathcal{L}_I$) and cross-modal projection ($\mathcal{L}_M$) losses result to a rank-1 accuracy of 49.85% when used together and significantly less when used individually. By introducing BERT we demonstrate that better word embeddings can be learned that increase the accuracy to 52.97%. Finally, when the proposed adversarial representation learning paradigm (ARL) is introduced, additional improvements are observed, regardless of whether BERT is used or not. We observe relative improvements of 2.9% and 3% with and without the utilization of BERT respectively, which demonstrates that ARL helps the network learn domain-invariant representations that

can successfully be used at deployment-time to perform cross-modal matching. Similar results are obtained in the Flickr30K dataset in which ARL improved the rank-1 matching performance from 51.2% to 53.1% and from 41.0% to 42.6% in image-to-text and text-to-image, respectively.

Impact Backbone Depth: In the second ablation study, we investigate to what extent the depth of the backbone networks affects the final cross-modal matching performance. Similar to previous well-performing methods [11, 63, 138], a fully-connected layer was used to learn the word embeddings (denoted by FC-Embed.) and compared it with the deep language model of BERT. For the image modality, two different ResNet backbones were employed while the rest of our proposed methodology remained the same. Rank-1 matching results in both directions are reported on the Flickr30K dataset in Table 5.4. Introducing a language model yields significant improvements (4.8% and 4.7%) regardless of the image backbone. In addition, increasing the image backbone depth results in smaller text-to-image matching improvements of approximately 2%.

Table 5.4: Ablation studies on the Flickr30K dataset to assess the impact of the depth of different backbone architectures on the final performance.

Image Backbone		Text Backbone		Img2Txt	Txt2Img
ResNet-101	ResNet-152	FC-Emb.	BERT	Rank-1	Rank-1
✓		✓		47.9	35.8
✓			✓	52.0	40.6
	✓	✓		50.1	37.9
	✓		✓	53.1	42.6

Impact of normalization in the distance: While the cosine distance focuses on the angle between vectors (thus not taking into consideration their weight or magnitude), Euclidean distance is similar

to using a ruler to actually measure the distance. Cosine similarity is generally used as a metric for measuring distance when the magnitude of the vectors does not matter. This happens for example when working with text data represented by word counts. Aiming to investigate to what extent the cosine distance impacts the final performance an ablation study on the CUHK-PEDES dataset was conducted. The cosine and the Euclidean distances were used to measure the distance between the probe textual features and the gallery image features and the obtained results are presented in Table 5.5. The cosine distance results into superior rank-1 accuracy compared to the Euclidean distance by 2.64%. The cosine distance has a special property that makes it suitable for metric learning: the resulting similarity measure is always within the range of $[-1, 1]$. This property is the reason why it is also frequently used in face recognition applications where experimental comparisons have demonstrated that it outperforms other alternatives [130].

Table 5.5: Ablation study on the impact of cosine distance on the CUHK-PEDES dataset.

Distance	**Rank-1**
Cosine	54.51
Euclidean	51.87

Qualitative Results: In Figure 5.2, cross-modal retrieval results for all four datasets are provided. TIDAM is capable of learning cloth and accessory-related correspondences as it can accurately retrieve images of people carrying bags with the correct set of clothing. In addition, TIDAM retrieves consistent images given a textual query (e.g., group of people in snow) as well as similar textual descriptions, given an image query (e.g., all three descriptions describe dogs or soccer players in the first and second row on the right).

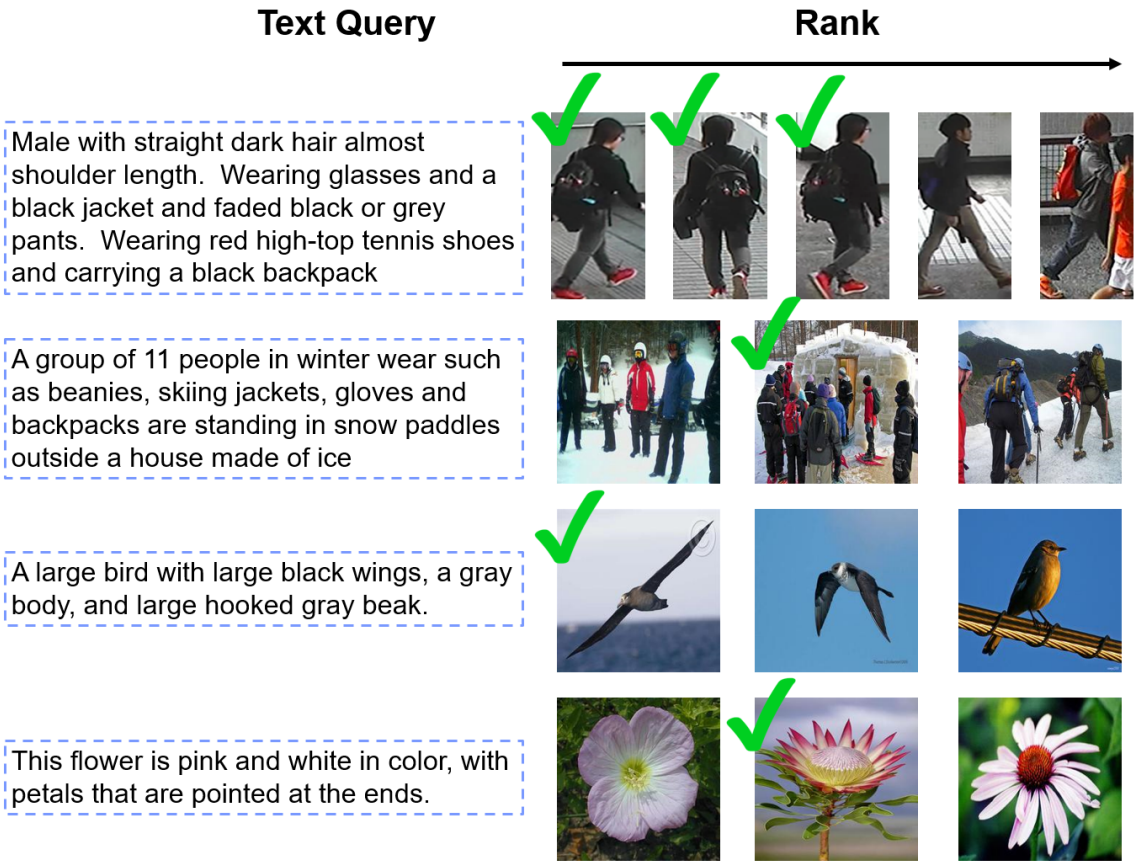

Figure 5.2: Given a textual description as a query, the most relevant images ranked from left to right are retrieved. Successful retrieval is performed in cases with poor lighting, under different poses, and with different visual attributes.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This dissertation is focused on the problem of visual attribute classification and person search. An ontology was first defined to perform such tasks and several algorithms were introduced that predict the visual attributes of humans as well as effectively perform person search.

An attribute ontology was proposed which comprises identity and identity-related attributes. Traits were extracted from the textual information and were then mapped to the attribute ontology that generates positive or negative labels for the set of attributes that it includes. For example, if the description contains words such as “man”, “guy”, “boy” then the *“Sex”* attribute of the ontology has a positive label. By using this ontology attribute pseudo-labels are extracted that can then be used to train attribute classification models with superior performance than previous work.

A series of methods were also introduced to tackle the visual attribute classification problem. Two novel techniques were introduced that employ the LUPI framework to predict soft-biometric

traits such as the gender or the height of an individual. Following that two deep-learning methods were designed that perform visual attribute classification in images of humans. Detailed experimental results, ablation studies and qualitative results were provided in each case to demonstrate the performance of the proposed approaches as well as the contribution of individual components. State-of-the-art-results were obtained in several publicly available datasets and further limitations were provided to facilitate future research.

Finally, a text-to-image matching approach was designed that can be applied to text-to-image person search applications. A domain discriminator was introduced that can help the network learn better feature representations from both the image and the textual inputs. In addition, it was demonstrated that deep language models are well-suited for this application as they can improve the retrieval performance. Experimental evaluations were performed in four widely-used publicly available datasets and state-of-the-art results were obtained across the board which demonstrates the efficacy of the proposed approach.

6.2 Future Work

Objective 1: Attribute Ontology: The ontology introduced in this dissertation covers attributes and modifiers based on the UHPD textual descriptions of the past five years as well as the most frequent attributes that appeared in the descriptions of the CUHK-PEDES dataset. The limitations along with the future work for this objective are the following:

1. The proposed ontology is limited to the most frequently used attributes. Thus, the present ontology could be expanded to include a wider variety of attributes and more transient attributes. For example, “tattoos” is an attribute that should be added to the ontology since it appears three times in the descriptions collected from the UHPD between 2013 and 2018.

2. The existing textual descriptions are insufficient in terms of the variety of attributes they cover and at the same time contain several mistakes. For example, the input sentences contain typos and mistakes, words are written in American as well as British English (e.g., gray and grey) and sentences demonstrate a large variance in what is described even for the same image. To address this limitation, a better tool for annotation could be developed that would ask the annotators to write detailed textual descriptions. Finally, careful data cleaning that would remove or correct erroneous textual queries would help our algorithms learn better representations and achieve better matching since the existing datasets contain several mistakes.

Objective 2: Visual Attributes: The algorithms designed in this dissertation predict the visual attributes of an individual given an image as an input. The limitations along with the future work for this objective are the following:

1. A significant limitation of most pedestrian attribute classification methods is that they resize the input data to a fixed square-size resolution (e.g., 224×224) in order to feed them to deep pre-trained architectures. Human crops are usually rectangular captured from different viewpoints and thus, when they are resized to a square, important spatial information is lost. To address this limitation, future work will need to feed the whole image (before performing the human crop) at a fixed resolution that does not destroy the spatial relations and then extract human-related features using ROI-pooling at a stage within the network.

2. A second source of error is the very low resolution of several images especially in the PETA dataset, which makes it hard even for the human eye to identify the attribute traits of the depicted human. In addition, the provided annotations contain a third unspecified/uncertain class, which is used as negative during training in the literature, that further dilutes the learning process. Applying modern super-resolution techniques [18, 53] could alleviate this issue

but only to some extent. Regarding errors due to modeling richer feature representations could be extracted using feature pyramid networks [69] since they extract high-level semantic feature maps at multiple scales.

3. The attention mechanism was designed with simplicity and speed in mind and thus it does not work as well as other alternatives in the literature. Modern visual attention mechanisms [66, 68, 119] could be adapted to a multi-label setup and applied to achieve superior performance at the expense of a larger parameter space.

Objective 3: Person Search: While the algorithms introduced in this thesis achieved state-of-the-art results in four publicly available widely-used datasets there is still room for improvement since the rank-1 accuracy is 55%. The limitations along with the future work for this objective are the following:

1. The relation between different attributes in the image is not explored and the approaches introduced in this dissertation treat each attribute separately without taking into consideration the connection between them. To learn better feature representations that can effectively perform cross-modal matching, recent graph convolutional neural network approaches could be utilized. After performing object and pedestrian detection in an image, the identified region proposal could be fed to such networks that can extract discriminative representations by taking into consideration the relation between the objects in the image.

2. The failure cases indicated that there are several examples in which the retrieved images completely match the textual descriptions but they are marked as incorrect since the retrieved identity is not the correct one. To address this limitation, the evaluation protocol would need to be expanded with a metric that demonstrates to what extent the retrieved image contains the attributes provided in the input textual queries.

Bibliography

[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*, pages 1–6, Seattle, WA, 2010.

[2] M. Andersen, J. Dahl, and L. Vandenberghe. CVXOPT free software package for convex optimization based on the Python programming language. (available online at http://cvxopt.org/).

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2425–2433, Boston, MA, 2015.

[4] Y. Bengio et al. Deep learning of representations for unsupervised and transfer learning. *ICML Unsupervised and Transfer Learning*, 27:17–36, 2012.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. International Conference on Machine Learning*, pages 41–48, Montreal, Canada, 2009.

[6] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Proc. IEEE International Conference on Computer Vision*, pages 1543–1550, Barcelona, Spain, 2011.

[7] D. Cao, C. Chen, D. Adjeroh, and A. Ross. Predicting gender and weight from human metrology using a copula model. In *Proc. IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 162–169, Washington DC, USA, 2012.

[8] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proc. European Conference on Computer Vision*, pages 135–150, Munich, Germany, 2018.

[9] H.-S. Chang, E. Learned-Miller, and A. McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Proc. Neural Information Processing Systems*, pages 1002–1012, Long Beach, CA, 2017.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proc. European Conference on Computer Vision*, pages 54–70, Munich, Germany, 2018.

[12] S.-F. Chen, Y.-C. Chen, C.-K. Yeh, and Y.-C. F. Wang. Order-free rnn with visual attention for multi-label classification. In *AAAI Conference on Artificial Intelligence*, pages 185–194, New Orleans, LA, 2018.

[13] T. Chen, C. Xu, and J. Luo. Improving text-based person search by spatial matching and adaptive threshold. In *Proc. Winter Conference on Applications of Computer Vision*, pages 1879–1887, Lake Tahoe, NV, 2018.

[14] X. Chen and B. Bhanu. Soft biometrics integrated multi-target tracking. In *Proc.* 22^{nd} *International Conference on Pattern Recognition*, pages 4146–4151, Stockholm, Sweden, Aug. 24-28 2014.

[15] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu. How to train a GAN? Tips and tricks to make GANs work. github.com/soumith/ganhacks, 2016.

[16] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, Honolulu, HI, 2017.

[17] 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*, pages 1–8, Providence, RI, 2012.

[18] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. In *Proc. International Conference on Computer Vision*, pages 5439–5448, Venice, Italy, 2017.

[19] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[20] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM Multimedia*, pages 789–792, Orlando, FL, 2014.

[21] Q. Dong, S. Gong, and X. Zhu. Class rectification hard mining for imbalanced deep learning. In *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, Oct. 22-29 2017.

[22] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *Winter Conference on Applications of Computer Vision*, pages 520–529, Santa Rosa, CA, 2017.

[23] C. Drummond, R. C. Holte, et al. Class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Proc. Workshop on Learning from Imbalanced Datasets II*, volume 11, pages 1–8, Washington, DC, 2003.

[24] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, Miami, FL, 2009.

[25] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proc. Advances in Neural Information Processing Systems*, pages 433–440, Vancouver, Canada, 2007.

[26] C. Florensa, D. Held, M. Wulfmeier, and P. Abbeel. Reverse curriculum generation for reinforcement learning. In *Proc. 1st Conference on Robot Learning*, pages 1–8, Mountain View, CA, 2017.

[27] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. International Conference on Machine Learning*, pages 1180–1189, Lille, France, 2015.

[28] R. Girshick. Fast R-CNN. In *Proc. International Conference on Computer Vision*, pages 1440–1448, Santiago, Chile, 2015.

[29] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *Proc. IEEE International Conference on Computer Vision*, pages 2470–2478, Santiago, Chile, 2015.

[30] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with R*CNN. In *Proc. IEEE International Conference on Computer Vision*, pages 1080–1088, Santiago, Chile, 2015.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems*, pages 2672–2680, Montréal Canada, 2014.

[32] 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*, pages 1–7, Rio, Brazil, 2007.

[33] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, Salt Lake City, UT, 2018.

[34] H. Guo, X. Fan, and S. Wang. Human attribute recognition by refining attention heat map. *Pattern Recognition Letters*, 94:38–45, 2017.

[35] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1–8, New Orleans, LA, 2017.

[36] B. Hariharan, L. Zelnik-Manor, M. Varma, and S. Vishwanathan. Large scale max-margin multi-label classification with priors. In *Proc. International Conference on Machine Learning*, pages 423–430, Haifa, Israel, 2010.

[37] 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*, pages 770–778, Las Vegas, NV, 2016.

[38] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[39] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, Las Vegas, NV, 2016.

[40] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, Honolulu, HI, 2017.

[41] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proc. IEEE International Conference on Computer Vision*, pages 1062–1070, Boston, MA, 2015.

[42] Q. Huang, W. Liu, and D. Lin. Person search in videos with one portrait through visual and temporal links. In *Proc. European Conference on Computer Vision*, pages 425–441, Munich, Germany, 2018.

[43] S.-J. Huang, Z.-H. Zhou, and Z. Zhou. Multi-label learning by exploiting label correlations locally. In *Proc. AAAI*, pages 1–8, Toronto, Canada, 2012.

[44] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, Lille, France, 2015.

[45] M. R. Islam, F. K.-S. Chan, and A. W.-K. Kong. A preliminary study of lower leg geometry as a soft biometric trait for forensic investigation. In *Proc. International Conference on Pattern Recognition*, pages 427–431, Stockholm, Sweden, 2014.

[46] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Proc. Neural Information Processing Systems*, pages 2017–2025, Montreal, Canada, 2015.

[47] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proc. International Conference on Biometric Authentication*, pages 731–738, Hong Kong, China, 2004.

[48] 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*, pages 1629–1636, Columbus, OH, 2014.

[49] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D video for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[50] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proc. International Conference on Machine Learning*, pages 1–8, Stockholm, Sweden, 2018.

[51] I. A. Kakadiaris, N. Sarafianos, and C. Nikou. Show me your body: Gender classification from still images. In *Proc. IEEE International Conference on Image Processing*, Phoenix, AZ, Sep. 25-28 2016.

[52] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2017.

[53] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, Las Vegas, NV, 2016.

[54] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations*, pages 1–8, San Diego, CA, 2015.

[55] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, Boston, MA, 2015.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in neural information processing systems*, pages 1097–1105, Lake Tahoe, 2012.

[57] 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.

[58] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[59] X. Lan, X. Zhu, and S. Gong. Person search by multi-scale matching. In *Proc. European Conference on Computer Vision*, pages 536–552, Munich, Germany, 2018.

[60] M. Lapin, M. Hein, and B. Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, 53:95–108, 2014.

[61] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proc. Asian Conference on Pattern Recognition*, pages 111–115, Nanjing, China, 2015.

[62] K. Li, J. Xing, W. Hu, and S. J. Maybank. D2c: Deep cumulatively and comparatively learning for human age estimation. *Pattern Recognition*, 66:95–105, 2017.

[63] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *Proc. International Conference on Computer Vision*, pages 1890–1899, Venice, Italy, 2017.

[64] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1970–1979, Honolulu, HI, 2017.

[65] S. Li, X. Zhu, Q. Huang, H. Xu, and C.-C. J. Kuo. Multiple instance curriculum learning for weakly supervised object detection. In *Proc. British Machine Vision Conference*, pages 1–8, London, UK, 2017.

[66] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, Salt Lake City, UT, 2018.

[67] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *Proc. European Conference on Computer Vision*, pages 684–700, Amsterdam, The Netherlands, 2016.

[68] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann. Focal visual-text attention for visual question answering. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 6135–6143, Salt Lake City, UT, 2018.

[69] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, Honolulu, HI, 2017.

[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proc. International Conference on Computer Vision*, pages 2980–2988, Venice, Italy, 2017.

[71] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar. Microsoft COCO: Common objects in context. In *Proc. European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland, 2014.

[72] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 212–220, Honolulu, HI, 2017.

[73] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision*, pages 3730–3738, Santiago, Chile, 2015.

[74] M. Livne, L. Sigal, N. F. Troje, and D. J. Fleet. Human attributes from 3d pose tracking. *Computer Vision and Image Understanding*, 116(5):648–660, 2012.

[75] M. Long, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *Proc. International Conference on Machine Learning*, pages 2208–2217, New York, NY, 2016.

[76] D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik. Unifying distillation and privileged information. In *Proc. International Conference on Learning Representations*, pages 1–8, San Jose, Puerto Rico, 2016.

[77] T. Maciejewski and J. Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *Proc. Computational Intelligence and Data Mining*, pages 104–111, Paris, France, 2011.

[78] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008.

[79] 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*, pages 1–6, Miyagi, Japan, 2016.

[80] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman. Teacher-student curriculum learning. In *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6-13 2017.

[81] E. Meyerson and R. Miikkulainen. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. In *Proc. International Conference on Learning Representations*, pages 1–8, Vancouver, Canada, 2018.

[82] A. Murali, L. Pinto, D. Gandhi, and A. Gupta. CASSL: Curriculum accelerated self-supervised learning. In *Proc. International Conference on Robotics and Automation*, pages 6453–6460, Mountain View, CA, 2018. IEEE.

[83] A. Nagrani, S. Albanie, and A. Zisserman. Learnable PINs: Cross-modal embeddings for person identity. In *Proc. European Conference on Computer Vision*, pages 71–88, Munich, Germany, 2018.

[84] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 299–307, Honolulu, HI, 2017.

[85] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):1–10, 2018.

[86] D. Parikh and K. Grauman. Relative attributes. In *Proc. IEEE International Conference on Computer Vision*, pages 503–510, Barcelona, Spain, 2011.

[87] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[88] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *Proc. European Conference on Computer Vision*, pages 38–56, Amsterdam, The Netherlands, 2016.

[89] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, Boston, MA, 2015.

[90] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Proc. Advances in neural information processing systems*, pages 557–563, 1999.

[91] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proc. International Conference on Computer Vision*, pages 2641–2649, Santiago, Chile, 2015.

[92] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 49–58, Las Vegas, NV, 2016.

[93] 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):211–252, 2015.

[94] SAE International. CAESAR: Civilian American and European Surface Anthropometry Resource database. (Available online at http://store.sae.org.caesar).

[95] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[96] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris. Curriculum learning of visual attribute clusters for multi-task classification. *Pattern Recognition*, 80:94–108, 2018.

[97] N. Sarafianos, M. Vrigkas, and I. A. Kakadiaris. Adaptive SVM+: Learning with privileged information for domain adaptation. In *Proc. International Conference on Computer Vision Workshops*, pages 2637–2644, Venice, Italy, 2017.

[98] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proc. European Conference on Computer Vision*, pages 680–697, Munich, Germany, 2018.

[99] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *Proc. British Machine Vision Conference*, pages 1–14, London, UK, 2017.

[100] 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*, pages 2933–2940, Providence, RI, 2012.

[101] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, Columbus, OH, 2014.

[102] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to transfer privileged information. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, Las Vegas, NV, 2014.

[103] Z. Shi, Y. Yang, T. Hospedales, and T. Xiang. Weakly-supervised image annotation and segmentation with objects and attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2525–2538, 2016.

[104] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations*, pages 1–14, San Diego, CA, 2015.

[105] A. J. Smola and B. Schölkopf. A tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, 2004.

[106] J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and X. Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *Proc. British Machine Vision Conference*, pages 1–14, York, UK, 2016.

[107] 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.

[108] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. In *Proc. International Conference on Machine Learning*, pages 4189–4198, Montreal, Canada, 2015.

[109] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*, 75:77–89, 2018.

[110] C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, and L. S. Davis. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition*, 66:4–15, 2017.

[111] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic CNN model. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 87–95, Boston, MA, 2015.

[112] 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*, pages 1139–1147, Atlanta, GA, 2013.

[113] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, MA, 2015.

[114] O. Tuzel, T. K. Marks, and S. Tambe. Robust face alignment using a mixture of invariant experts. In *Proc. European Conference on Computer Vision*, pages 825–841, Amsterdam, The Netherlands, 2016.

[115] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proc. International Conference on Computer Vision*, pages 4068–4076, Santiago, Chile, 2015.

[116] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–57, 2009.

[117] B. Victor, K. Bowyer, and S. Sarka. An evaluation of face and ear biometrics. In *Proc. International Conference on Pattern Recognition*, pages 429–432, Quebec, Canada, 2002.

[118] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, Boston, MA, 2015.

[119] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, Honolulu, HI, 2017.

[120] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: L2 hypersphere embedding for face verification. In *Proc. ACM on Multimedia Conference*, pages 1041–1049, Mountain View, CA, Oct. 23-27 2017.

[121] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. In *Winter Conference on Applications of Computer Vision*, pages 530–539, Santa Rosa, CA, 2017.

[122] S. Wang, D. Tao, and J. Yang. Relative attribute SVM+ learning for age estimation. *IEEE Transactions on Cybernetics*, 46(3):827–839, 2015.

[123] W. Wang and J. Shen. Deep visual attention prediction. *Transactions on Image Processing*, 27(5):2368–2378, 2018.

[124] W. Wang, Y. Yan, S. Winkler, and N. Sebe. Category specific dictionary learning for attribute specific feature selection. *IEEE Transactions on Image Processing*, 25(3):1465–1478, 2016.

[125] J. H. J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[126] G. Williams, G. Taylor, K. Smolskiy, and C. Bregler. Body motion analysis for multi-modal identity verification. In *Proc. International Conference on Pattern Recognition*, pages 2198–2201, Istanbul, Turkey, 2010.

[127] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng. Ian: the individual aggregation network for person search. *Pattern Recognition*, 87:332–340, 2019.

[128] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, Honolulu, HI, 2017.

[129] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66:106–116, 2017.

[130] X. Xu, H. A. Le, P. Dou, Y. Wu, and I. A. Kakadiaris. Evaluation of a 3d-aided pose invariant 2d face recognition system. In *Proc. IEEE International Joint Conference on Biometrics*, pages 446–455, 2017.

[131] X. Xu, W. Li, and D. Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(12):3150–3162, 2015.

[132] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proc. 15th ACM International Conference on Multimedia*, pages 188–197, 2007.

[133] 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*, pages 3320–3328, Montreal, Canada, 2014.

[134] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[135] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, Las Vegas, NV, 2016.

[136] 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*, pages 2718–2726, Las Vegas, NV, 2016.

[137] 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*, pages 1637–1644, Columbus, OH, 2014.

[138] Y. Zhang and H. Lu. Deep cross-modal projection learning for image-text matching. In *Proc. European Conference on Computer Vision*, pages 686–701, Munich, Germany, 2018.

[139] J. Zheng, Z. Jiang, and R. Chellappa. Submodular attribute selection for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2242–2255, 2016.

[140] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen. Dual-path convolutional image-text embedding with instance loss. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 212–220, Honolulu, HI, 2017.

[141] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2425–2433, Boston, MA, 2015.

[142] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, Honolulu, HI, 2017.

[143] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229, 2016.

[144] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *Proc. International Conference on Biometrics*, pages 535–540, Phuket, Thailand, 2015.