Deep View-Sensitive Pedestrian Attribute Inference in an end-to-end Model

M. Saquib Sarfraz*[1]
saquib.sarfraz@kit.edu
Arne Schumann*[2]
arne.schumann@iosb.fraunhofer.de
Yan Wang[1]
uudqa@student.kit.edu
Rainer Stiefelhagen[1]
rainer.stiefelhagen@kit.edu

[1] Institute of Anthropomatics & Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2] Fraunhofer IOSB
Fraunhoferstr. 1,
Karlsruhe, Germany

Abstract

Pedestrian attribute inference is a demanding problem in visual surveillance that can facilitate person retrieval, search and indexing. To exploit semantic relations between attributes, recent research treats it as a multi-label image classification task. The visual cues hinting at attributes can be strongly localized and inference of person attributes such as hair, backpack, shorts, etc., are highly dependent on the acquired view of the pedestrian. In this paper we assert this dependence in an end-to-end learning framework and show that a view-sensitive attribute inference is able to learn better attribute predictions. Our proposed model jointly predicts the coarse pose (view) of the pedestrian and learns specialized view-specific multi-label attribute predictions. We show in an extensive evaluation on three challenging datasets (PETA, RAP and WIDER) that our proposed end-to-end view-aware attribute prediction model provides competitive performance and improves on the published state-of-the-art on these datasets.

1 Introduction

Person attribute recognition in surveillance footage is a highly demanding problem as it benefits several related applications such as image indexing, person retrieval [7, 18] and person re-identification [10]. Methods addressing pedestrian attribute recognition in such applications have to deal with challenges due to low resolution, detected pedestrians in far-range surveillance scenes, pose variations, and occlusions. The task is to make predictions for a set of attributes given an image of a person as input. In contrast to the general image recognition problem where each image has one label of a certain class, the pedestrian attribute inference is a multi-label recognition problem where each of the pedestrian images is assigned a multitude of semantic attribute labels with binary outcome, e.g., wearing short skirt, male, running, carrying backpack, etc.

*Arne Schumann and M. Saquib Sarfraz contributed equally to this work.

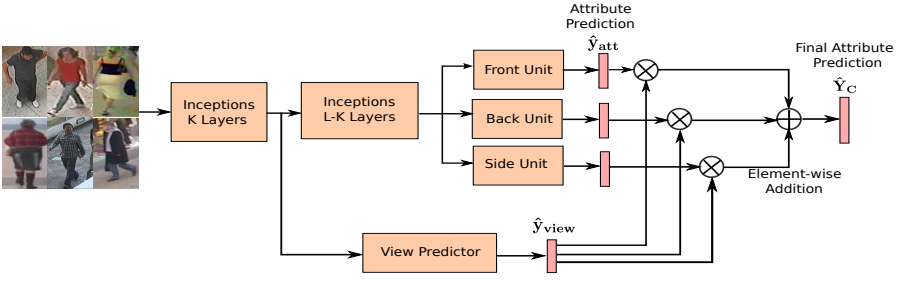

Figure 1: Overview of our View-Sensitive Pedestrian Attribute (VeSPA) model.

Most previous methods solve the multi-label person attribute recognition by optimizing a separate binary classification model for each of the attributes [11, 19, 24]. Such a direction, however, ignores the semantic relationships between attributes. To better exploit these semantic relationships more recent proposals have shown improved performance by solving the task as a direct multi-label classification problem [12, 13, 26]. While multi-label classification generally achieves better accuracy than binary classifiaction approaches, few approaches explicitly capture the spatial relationships of attributes (i.e. the location of attributes in the image). This can have several associated problems as each labeled attribute may intrinsically be tied to very different image regions based on the acquired view of the pedestrian. As an example, carrying backpack in a back-view image has a very different and much larger spatial context to learn from in comparison with a side- or front-view. If this spatial context is known, it may better guide the training process to focus on respective image regions for each attribute. In a very recent proposal Zhu et al. [26] try to infer this localization explicitly in the model and have shown improved performance. While spatial context is relevant for some of the attributes it is not pertinent for other, more global attributes such as gender, action, age, etc.

Another solution is to guide the learning process by using additional part or pose information. As in the example above, the spatial context is also tied to the pose of the person. The pose information may also provide better context in explaining some of the global level attributes like gender, age, or action. Some recent methods [24] [13] proposed to learn separate models depending upon *a priori* selected parts, pose and context. In this paper we propose to learn an end-to-end unified model to jointly learn the coarse pose prediction (front-, back-, or side-view) and specialized, view-dependent attribute inference in a multi-label classification setting. We use a deep Convolutional Neural Network (CNN) to train a view predictor in the early layers and separate view-specific attribute prediction units in the later layers of the same model. Figure 1 depicts an overview of our View-Sensitive Pedestrian Attribute approach (VeSPA).

Our main contribution is to show that in addition to the popular view of relying on either body parts, attribute spatial context in the image, or general scene context, the coarse body pose (view) information can be another simple yet highly relevant clue for reliable attributes inference. Our results show that different attributes relate differently to the acquired view of the person and learning the views helps the overall attribute prediction more. Our evaluation on three of the largest available datasets, the PETA and RAP surveillance datasets, and the WIDER attribute dataset which features challenging person images in photos, shows

convincing results.

In the remainder of this work we first discuss related approaches in Section 2. Section 3 provides details of the proposed approach. We conclude in Section 5 after discussing the results of our comprehensive evaluations in Section 4.

2 Related Work

We limit our discussion of related work to attribute recognition approaches which pertain to person or pedestrian images.

Attribute classification is a multi-label classification task. A straightforward way to address this is by relying on the extensive single-label classification literature and training a separate classifier for each attribute. Sharma et al. [15] apply this approach by using spatial histogram features in conjunction with a maximum margin optimization to learn each of the attribute classes. Similarly, Layne et al. [10] and Deng et al. [4] use Support Vector Machines (SVMs) and a set of color and texture features to classify each attribute. Such approaches cannot directly leverage semantic relations between attributes. More recent approaches often rely on a single model to recognize all attributes (multi-label classification). Sudowe et al. [19] describe a person attribute CNN which is trained with one loss for each attribute. The main part of the net is shared for all attributes and allows the approach to implicitly leverage attribute relationship information. A similar approach with individual attribute losses which are manually restricted to relevant body parts is described by Zhu et al. in [25]. In [11] Li et al. show that it can be beneficial to train an attribute CNN with a single, weighted loss which includes all attributes and applies weights based on each attribute's label imbalance. This approach is extended by Li et al. in [12] into a part-based model. In [22] Yu et al. propose a CNN based approach which relies on multi-level deep features and is able to recognize as well as localize pedestrian attributes. This approach also relies on a single, weighted loss. Joint recognition and localization of general image attributes is also performed in [26] and applied to person attribute recognition.

The importance of person pose information for the attribute recognition task has been studied in several works. In [2] Bourdev et al. use pose-sensitive body part detectors and apply attribute classifiers for each part detection. Each attribute classifier is thus specific to a certain body part and pose. A main drawback is the large number of resulting attribute classifiers. In a later work [24] the same part detectors are used to generate a pose-normalized person representation based on deep features which is used for attribute recognition with linear SVMs. Park et al. [14] describe a deep model which jointly learns to detect person keypoints, body parts and attributes. Pose information is implicitly contained in normalized body part representation and attributes are manually assigned to the relevant body parts. In [21] Yang et al. learn a joint model for body part localization and attribute recognition which detects keypoints and generates a warping matrix for pose normalization. Another recent approach by Li et al. [13] relies on full image and leverages body parts and scene context information to more accurately determine person attributes in a combined deep model. Most of these approaches aim to include additional information, localization (explicitly or by a part-based approach) or context knowledge, to aid in the attribute recognition task. However, they do not explicitly rely on a person's acquired view which our experiments show is a crucial clue for robust attribute recognition.

3 View-Sensitive Pedestrian Attribute Inference

We adapt a deep neural network for joint pose and multi-label attribute classification. The overall design of our approach is shown in Figure 1. The main network is based on the GoogleNet inception architecture [20]. As shown, the network contains a view classification branch and three view-specific attribute predictor units. The view classifier and attribute predictors are both trained with separate loss functions. Prediction scores from weighted view-specific predictors are aggregated to generate the final multi-class attribute predictions. The whole network is a unified framework and is trained in an end-to-end manner.

3.1 End-to-end View & View-aware Attribute Classification

Let $\mathbf{I}$ denote an input person image with ground-truth labels $Y = \{Y_C; Y_V\}$. $Y_C = [y^1, y^2, ..., y^C]^T$ denotes the attribute labels, where y^i is a binary indicator: $y^i = 1$, if image $\mathbf{I}$ is tagged with attribute i and $y^i = 0$ otherwise. C is the number of attributes in the dataset. The view label is denoted by $Y_V \in \{front, back, side\}$. The overall network conducts view prediction and multi-label attribute inference using two different corresponding losses.

The main net is based on the GoogleNet inception architecture. It has repetitive inception blocks where each inception module ranges from 256 filters in the early modules to 1024 in top inception modules. Our design shares the same network for both tasks. The coarse pose or view prediction is based on the output of early layers. The view predictor is a branch out after the K-th layer of the model and conducts classification for each of the three views Y_V of the image $\mathbf{I}$:

$$\begin{aligned} \mathbf{X}_K &= f_K(\mathbf{I}; \theta_K), \;\; \mathbf{X}_K \in \mathbb{R}^{n \times n \times k} \\ \hat{\mathbf{y}}_{view} &= f_{view}(\mathbf{X}_K; \theta_{view}), \;\; \hat{\mathbf{y}}_{view} \in \mathbb{R}^3 \end{aligned} \tag{1}$$

Here, $\mathbf{X}_K$ are the k feature maps of size $n \times n$ from layer K and $\hat{\mathbf{y}}_{view} = [\hat{y}^1_{view}, \cdots, \hat{y}^V_{view}]^T$ are the confidences of the view-predictor. These confidences act as weights for the output of the corresponding view-specific attribute inference units.

The attribute inference is carried out by branching out separate CNN units, one for each view. The input to each view-specific unit is the output feature map $\mathbf{X}_{L-1}$ of top level layer $L-1$. The output of each view-specific unit is an attribute prediction $\hat{\mathbf{y}}_{att} = [y^1, y^2, ..., y^C]^T$ for all the C attributes classes:

$$\begin{aligned} \mathbf{X}_{L-1} &= f_{L-1}(\mathbf{I}; \theta_{L-1}), \;\; \mathbf{X}_{L-1} \in \mathbb{R}^{n \times n \times k} \\ \hat{\mathbf{y}}_{att} &= f_{att}(\mathbf{X}_{L-1}; \Theta), \;\; \hat{\mathbf{y}}_{att} \in \mathbb{R}^C \end{aligned} \tag{2}$$

Here, $\Theta = \{\theta_{L-1}; \theta_{view}\}$ are the model parameters of the whole net. The view-specific attribute predictions are weighted by the corresponding predicted view-confidence $\hat{\mathbf{y}}_{view}$. The final multi-class attribute prediction $\hat{\mathbf{Y}}_\mathbf{C} = [y^1, y^2, ..., y^C]^T$ are the aggregated predictions of these weighted view-specific predictions:

$$\hat{\mathbf{Y}}_\mathbf{C} = \sum_V \hat{\mathbf{y}}^V_{att} \circ \hat{\mathbf{y}}^V_{view} \tag{3}$$

Note that the weighting by view predictor confidence not only weights the output of the attribute unit but importantly also weights its gradient equivalently. Thus, a unit whose corresponding view prediction weight is low will only receive very small parameter updates for the current training sample and thus focus its learning mainly on those samples that receive a correspondingly high confidence for the given view. With this design each of the three view-specific units is specialized in inferring attributes from the respective view and the final aggregation helps share the strengths of each of the units to improve the final prediction on all images. The whole network is trained with two losses. We use a 3-way softmax for the view-predictor branch. To cast the problem as a multi-label classification, following [11], we use a modified weighted cross-entropy loss at the final attribute inference layer:

$$L_{attr} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c(y_{ic})log(\hat{y}_{ic}) + (1-y_{ic})log(1-\hat{y}_{ic}) \tag{4}$$

where $w_c = exp(-a_c)$ is the weight for c-th attribute. a_c is the prior distribution of the c-th attribute in the training set. This is important, because of the large imbalance of attribute label values in the dataset. $\hat{y}_{ic}$ is the estimated probability for the c-th attribute of the image $\mathbf{I}$. During training we prevent the gradient of the multi-label attribute from flowing back through the pose predictor branch. Both, the view predictor gradient and attribute gradient flow back through the earlier K-layers which are updated by the sum of these two gradients.

Implementation Details: The standard GoogleNet inception architecture has two auxiliary loss layers that serve to strengthen the gradient and encourage discrimination in the lower stages of the network. In our design, we adapt one of these auxiliary loss layers as our view predictor and remove the other. The view predictor takes as input the 576 feature maps from the *inception 3c* layer of the net and comprises of a 5x5 max pooling, 1x1 conv block and two fully connected *fc* layers, where the last *fc* is the 3-dimensional input to the softmax to predict the view of the input pedestrian image. Each of the view-specific attribute units is a separate inception block which takes as input the 1024 feature maps of the model's *inception 5a* layer. Each unit contains 4 parallel strands which use chained convolutional blocks or pooling to achieve views of varying receptive fields on the input data. The output of each unit is concatenated, pooled and fed into a *fc* layer of C dimensions as the final output of the unit. The output of all three units are weighted by the view-predictor by multiplication and aggregated together by element-wise addition before being finally fed to the attribute loss.

To avoid overfitting and achieve a more robust attribute recognition, we apply image augmentation during training. Images are first normalized to zero-mean and resized to 256×256. For batch creation we then randomly crop the images to the GoogleNet input size of 227×227 and apply random horizontal flipping. The network's weights are initialized from a pre-trained ImageNet model and fine-tuned with an initial learning rate of 0.0002 using the Adam solver with a batch size of 32.

4 Evaluation

We evaluate our VeSPA model on three public datasets. The **PETA dataset** [3] is a collection of several person surveillance datasets and consists of 19,000 cropped images. Each image is annotated with 61 binary and 5 multi-value attributes. Following the established protocol, we limit our experiments to those 35 attributes for which the ratio of positive labels is higher

Method	RAP					PETA				
	mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1
ACN [19]	69.66	62.61	**80.12**	72.26	75.98	81.15	73.66	84.06	81.26	82.64
DeepMAR [11]	73.79	62.02	74.92	76.21	75.56	82.89	75.07	83.68	83.14	83.41
DeepMAR* [12]†	74.44	63.67	76.53	77.47	77.00	-	-	-	-	-
WPAL-GMP [22]†	**81.25**	50.30	57.17	78.39	66.12	**85.50**	76.98	84.07	**85.78**	84.90
WPAL-FSPP [22]†	79.48	53.30	60.82	78.80	68.65	84.16	74.62	82.66	85.16	83.40
GoogleNet Baseline	70.11	60.88	76.62	72.65	74.58	81.98	76.06	84.78	83.97	84.37
Ours VeSPA	77.70	**67.35**	79.51	**79.67**	**79.59**	83.45	**77.73**	**86.18**	84.81	**85.49**

Table 1: Results of our approach on the RAP and PETA datasets. We outperform the state-of-the-art on most of the example based metrics. Our strong F1 score indicates a better tradeoff between precision and recall than other works (Unpublished works are marked with †).

than 5%. The dataset is sampled into 9,500 training images, 1,900 validation images and 7,600 test images. The **RAP dataset** [12] consists of 41,585 person images recorded by surveillance cameras. Each image is annotated with 72 attributes, viewpoints, occlusions and body parts. According to the official protocol only those 51 attributes with a positive label ratio above 1% are used. For our evaluations we split the dataset randomly into 33,268 training images and 8,317 test images. In order to get a better impression of the performance of our model on data with more complex pose variation, we also evaluate on the **WIDER dataset** [13]. The dataset contains 13,789 full images with 57,524 person bounding boxes. 14 attributes are annotated for each person. We follow the evaluation protocol proposed in [26] and crop out all bounding boxes. This results in 28,340 person images for training and validation and 29,177 images for testing. We use all 14 attributes for our experiments on WIDER. The *unspecified* labels of the WIDER dataset are treated as negative during training and are excluded from evaluation in testing following the settings in [13] [26].

Of these three datasets only RAP contains view annotations and allows us to train the view predictor part of VeSPA. For training on WIDER and PETA we transfer the model learned on the RAP training data and fix the learning rate of the view predictor at 0 while training the remainder of the network as usual.

For our evaluations on PETA and RAP we rely on two types of metrics. For a *label-based* evaluation we compute the mean accuracy (mA) as the mean of the accuracy among positive examples and the accuracy among negative examples of an attribute. This metric is not affected by class imbalances and thus penalizes errors made for the less and more frequent label value equally strongly. However, this metric does not account for attribute relationships (i.e. consistency among attributes for a given person example). In order to account for attribute predictions which are consistent within each person image, we further use *example-based* metrics. For this we apply the well known metrics accuracy, precision, recall and F1 score averaged across all examples in the test data. A more detailed description of the metrics can be found in [12].

4.1 Comparison with State-of-the-art

We compare the performance of our VeSPA model with a number of recent state-of-the-art pedestrian attribute recognition works, including ACN [19], DeepMAR [11], DeepMAR* [12] and WPAL [22]. Results of our approach in context of these works are given in Table 1. We also include results of our GoogleNet baseline (without view-units) to demonstrate the gain of the proposed architecture. As seen, the model with view-units performs better. This

Method	RCNN [8]	R*CNN [9]	DHC [13]	ResNet-SRN [26]	ours VeSPA
mAP	80.0	80.5	81.3	86.2	82.4

Table 2: Results of our approach on the WIDER dataset. VeSPA achieves competitive performance in spite of the much stronger pose variation on this dataset.

benefit is more clearly demonstrated on the larger (with more attributes) RAP dataset, where a gain of 4-7% is achieved across all metrics, with using the specialized view-units. The gain is less pronounced on the PETA dataset (on average 1-2%).

Our approach achieves competitive performance across all metrics and state-of-the-art results on some of them. We strongly outperform most other approaches on the example based metrics. Particularly on accuracy and F1 our approach yields notable improvements over the previous state-of-the-art. The strong F1 values indicate a good precision-recall tradeoff of our approach. On both datasets the label based mean accuracy (mA) is lower than that of the unpublished WPAL approach. However, the example based results of WPAL are much lower in comparison. We have observed a similar trend during training of our approach. Prolonged training with possible overfitting will increase the mA metric further at the cost of all example based metrics. We argue that the example based metrics are more relevant to real world applications as they measure the consistency of an attribute-based description of a person which is of greater importance for communicating such descriptions to security personnel. Furthermore, description consistency is also important for subsequent tasks, such as person re-identification.

Both, PETA and RAP are typical surveillance datasets (i.e. pedestrian attribute recognition). While they offer great variety in view angle, they do not contain a very high degree of person body pose variation. In order to judge the accuracy of VeSPA under stronger/unknown pose variations, we further compare our performance to the state-of-the-art on the WIDER dataset (i.e. person attribute recognition). We compare our performance to R-CNN [8], R*CNN [9], DHC [13] and the recent ResNet-SRN [26]. Results are shown in Table 2. Our approach outperforms the published state-of-the-art, i.e. R-CNN, R*CNN and DHC by at least 1.1% in mean average precision (mAP). Interestingly, two of these methods (R*CNN and DHC) make explicit use of scene context and image parts to increase person attribute precision. Our results show that view information might be a more relevant clue for attribute recognition than context. The most recent approach, ResNet-SRN [26] which simultaneously recognizes and localizes attributes and uses the image-level context of each attribute outperforms VeSPA in mAP. However, the ResNet-SRN model is highly optimized to learn the image-level context cues from the WIDER training set whereas our view prediction is merely transfered from the RAP dataset and no adaptation to views or poses present in WIDER is learned.

Our approach shows competitive or state-of-the-art performance across the three datasets. Compared to the previous published state-of-the-art, VeSPA has a superior precision-recall trade-off which makes it particularly suitable for applications that rely on an accurate and consistent attribute-based description of persons.

4.2 View Prediction Analysis

We investigate the effect of view prediction in aiding the overall attribute inference. We first provide some quantitative and qualitative performance of the view-classification on the three

Pass through	mA	mA	mA
Front unit	**80.02**	75.70	74.26
Back unit	73.40	**78.73**	73.92
Side unit	78.39	77.96	**76.73**

Figure 2: Specialization of each view-unit: Attributes mA is calculated by passing the subset of test images belonging to one of the three views through each of the three individual view units, respectively. The bold numbers denote the best performances that are obtained when the images of a specific view are processed by the matching view unit.

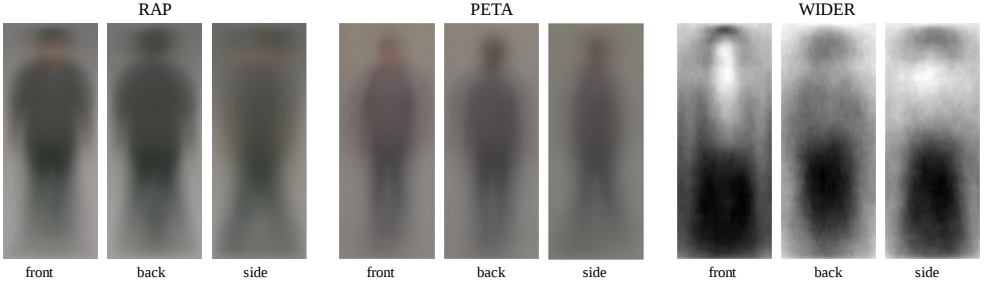

Figure 3: Mean images of VeSPAs view classification on the test sets of RAP, PETA and WIDER. **PETA:** Total test images=7600; predicted front (2414), back (2481), side (2705). **WIDER:** Total test images=29177; predicted front (7216), back (10774), side (11187). Best viewed in color and on screen

datasets. RAP is the only dataset which contains view annotations and thus allows for a quantitative evaluation. Our model achieves a very reliable view classification accuracy of 91.7%, 91.0% and 81.3% for front-view, back-view and side-view, respectively on RAP test set . This high accuracy is crucial to our approach as it allows the specialized attribute units of our model to reliably learn view specific information.

In order to quantitatively analyze if the trained view units are indeed specialized to the respective view, we divide the RAP test set according to the three view annotations and tested images of each view separately with all three view-units. For each test, the angle predictor and the other two view-units in the VeSPA model are deactivated. The results are displayed in Figure 2. The resulting attribute classification accuracy is always the highest for the respective matching view-unit. This shows that our architecture leads to a successful specialization of the view-units.

To gain insight into the effects of transferring the view predictor when training VeSPA on datasets which do not contain view annotations, we provide a qualitative impression in

Figure 4: Regions most relevant to VeSPAs attribute predictions: Excitation backprop [23] is used to obtain the region localizations on which our model bases its respective attribute prediction. The original confidence of VeSPA for the given attribute is also plotted on the bottom of the images as a red bar and the view confidences as green, yellow, and blue bars for front, back, and side, respectively. Each row of images corresponds to a specific view. Each column shows some representative attribute images from one of the three datasets (RAP, PETA, WIDER). This figure is best viewed in color.

Figure 3 by computing the mean image for each predicted view across all images on the PETA and WIDER test sets. We compare these mean images to the RAP mean images which we know from our quantitative analysis to represent very accurate view predictions. The figure shows a very high resemblance between the PETA and RAP mean images. This indicates a similarly high view classification accuracy on PETA as on RAP. The three views can be clearly identified by looking at the images. The side view images are slightly more ambiguous, because left-side and right-side are not differentiated by our model. WIDER dataset consists of photos with a much higher degree of background variability, clutter and person pose. This, and a large number of test images (29177) than those of PETA and RAP, leads to an increased blur of the mean image. To make the WIDER image viewable we only show the luminance channel of the color mean image. As seen, in comparison between back-view and front-view a lighter facial region is still clearly discernible on WIDER.

We also analyze which image regions are the most relevant to VeSPA's prediction of a given attribute. To that end, the excitation back-propagation method proposed by [23] is applied to our model to generate the attention maps for different attributes. Some of the representative results are shown in Figure 4 [1]. The images show that, importantly, VeSPA has been able to successfully identify different relevant image regions for the same attribute, under varying views, even though no localization is explicitly included in the training process.

[1] See more analysis of such relevant image regions for additional attributes in the supplementary material.

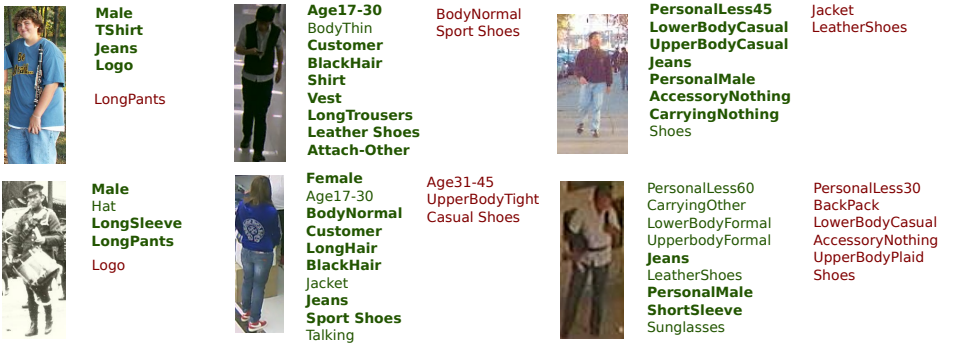

Figure 5: Qualitative results of VeSPA on WIDER (left), RAP (middle) and PETA (right). Correct attribute predictions are marked in bold green, missed attributes in green and false positive predictions in red. A negative example with a number of mistakes is diplayed in the lower right coner. Note however, that many of the mistakes are reasonable ones (e.g. predicting *shoes* instead of *leather shoes*).

For example, the relevant region for the presence of the attribute muffler is the neck area of a person in back-view images but the torso region in side-view images. Some interesting insights are on the clues that are considered most relevant for an attribute which has no well defined localized appearance in the image, e.g. the most relevant clue for the attribute *formal* appears to be the neckline region.

Our qualitative analysis shows that the model predictions are indeed based on meaningful attribute localizations and image context. Some example results of VeSPA are also depicted in Figure 5.

5 Conclusion

We have presented a unified model to jointly predict the person's view and specialized view dependent attribute inference. Our results shows that our model learns a reliable view predictor which is directly transferable to other datasets. The induced view-specific information into the attribute prediction units helps learn attributes better. In comparison to the published state-of-the-art that explicitly uses body parts, image context and scene context, our results show that relatively straight forward extensions and incorporating view information has proven useful for person attribute recognition. In addition to providing convincing semantic attribute predictions the view information may also aid in specific pedestrian search and retrieval applications.

References

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17, 2015.

[2] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-

based approach to attribute classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[3] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia (ACM-MM)*, 2014.

[4] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*, 2015.

[5] Ali Diba, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Deepcamp: Deep convolutional action & attribute mid-level patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Qi Dong, Shaogang Gong, and Xiatian Zhu. Multi-task curriculum transfer deep learning of clothing attributes. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[7] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 2014.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Person re-identification by attributes. In *British Machine Vision Conference (BMVC)*, 2012.

[11] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of the IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015.

[12] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[13] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[14] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human attributes, part and pose. *arXiv preprint arXiv:1605.02112*, 2016.

[15] Gaurav Sharma and Frederic Jurie. Learning discriminative spatial representation for image classification. In *British Machine Vision Conference (BMVC)*, 2011.

[16] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Expanded parts model for human attribute and action recognition in still images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[17] Jie Shen, Guangcan Liu, Jia Chen, Yuqiang Fang, Jianbin Xie, Yong Yu, and Shuicheng Yan. Unified structured learning for simultaneous human pose estimation and garment attribute classification. *IEEE Transactions on Image Processing*, 23, 2014.

[18] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015.

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] Luwei Yang, Ligen Zhu, Yichen Wei, Shuang Liang, and Ping Tan. Attribute recognition from adaptive parts. *arXiv preprint arXiv:1607.01437*, 2016.

[22] Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016.

[23] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016.

[24] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[25] Weipeng Zhang, Jie Shen, Guangcan Liu, and Yong Yu. A latent clothing attribute approach for human pose estimation. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[26] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *arXiv preprint arXiv:1702.05891, accepted CVPR*, 2017.

[27] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013.

[28] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *Proceedings of the International Conference on Biometrics (ICB)*, 2015.

[29] Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58, 2017.