Distraction-Aware Feature Learning for Human Attribute Recognition via Coarse-to-Fine Attention Mechanism

Mingda Wu,[1] Di Huang,[1]* Yuanfang Guo,[2] Yunhong Wang[1]
[1]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China
[2]IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing, China
{md99504, dhuang, andyguo, yhwang}@buaa.edu.cn

Abstract

Recently, Human Attribute Recognition (HAR) has become a hot topic due to its scientific challenges and application potentials, where localizing attributes is a crucial stage but not well handled. In this paper, we propose a novel deep learning approach to HAR, namely Distraction-aware HAR (DaHAR). It enhances deep CNN feature learning by improving attribute localization through a coarse-to-fine attention mechanism. At the coarse step, a self-mask block is built to roughly discriminate and reduce distractions, while at the fine step, a masked attention branch is applied to further eliminate irrelevant regions. Thanks to this mechanism, feature learning is more accurate, especially when heavy occlusions and complex backgrounds exist. Extensive experiments are conducted on the WIDER-Attribute and RAP databases, and state-of-the-art results are achieved, demonstrating the effectiveness of the proposed approach.

Introduction

Given an input image with a target person, Human Attribute Recognition (HAR) predicts his or her semantic characteristics, including low-level ones (*e.g.* wearing logo or plaid), mid-level ones (*e.g.* wearing hat or T-shirt), and high-level ones (*e.g.* gender, dressing formally). Accurate recognition of such attributes not only improves machine intelligence on cognition of humans, but also benefits a large number of applications such as person re-identification (Ling et al. 2019; Han et al. 2018), pedestrian detection (Tian et al. 2015), and person retrieval (Feris et al. 2014; Wang et al. 2013).

Existing investigations of HAR can be classified into three domains, *i.e.*, clothing domain, surveillance domain, and general domain. The techniques in the clothing domain have received extensive attentions (Al-Halah, Stiefelhagen, and Grauman 2017; Sarafianos, Vrigkas, and Kakadiaris 2017; Liu et al. 2016; Chen et al. 2015) due to their potentials in commercial applications. This type of methods generally require the input images of high resolutions with persons at a small number of pre-defined poses, and fine-grained clothing style recognition is still challenging. There are also numerous studies in the surveillance domain (Wang et al. 2019; Gao et al. 2019; Li et al. 2019; Liu et al. 2018; Wang et al. 2017), because these techniques are playing an important role in public security. Input images are recorded by a diversity of monitoring cameras, and the major difficulties lie in low resolutions, high blurrinesses, and complex backgrounds. In the past several years, interests have been shown in the general domain (Guo, Fan, and Wang 2017; Sarfraz et al. 2017; Li et al. 2016b), where input images are acquired in arbitrary scenarios exhibiting additional variations of gestures, viewpoints, illuminations, and occlusions, as depicted in Figure 1.

Figure 1: Human Attribute Recognition is challenging due to large variations of body gestures, external occlusions, lighting conditions, image resolutions and blurrinesses.

Regardless of differences between domains, HAR methods basically share a common framework, which conducts attribute-sensitive feature extraction on attribute-related regions for classification. In the literature, the majority of existing efforts to HAR have been made on building effective features, and a large number of works focus on improving the discrimination and the robustness of representations of appearance properties. Features are evolving from handcrafted ones (Joo, Wang, and Zhu 2013; Cao et al. 2008) to deep learned ones (Zhu et al. 2017b; Yu et al. 2016), with promising performance achieved.

To generate qualified features, attribute-related region localization is very crucial, which aims to locate the regions that contain useful clues for attribute recognition. With incorrect localization of such regions, attribute prediction tends to fail because meaningful features can hardly be captured. In both the surveillance and general domains (especially the latter one), the accuracy of attribute localization is usually susceptible to pose variations and external occlusions. Therefore, recent attempts develop more sophisticated schemes to locate attribute-related regions, including ex-

*Corresponding author: Di Huang.

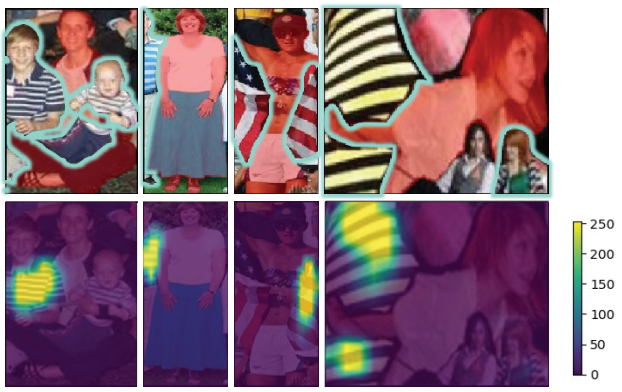

Figure 2: Examples of distractions (best scene in color). *Input images (1st row)* in HAR usually contain target subjects (masked in red) and distractions (enclosed in green contours). We visualize some cases when the attention mechanism mistakenly responds to *surrounding people (1st and 2nd columns)*, *similar objects (3rd column)* and *puzzling backgrounds (4th column)*, leading to *incorrect outputs (2nd row)*.

ploiting auxiliary cues of target persons (Fabbri, Calderara, and Cucchiara 2017; Li et al. 2018; Park, Nie, and Zhu 2017; Tan et al. 2019) and applying the attention mechanism (Tan et al. 2019; Sarafianos, Xu, and Kakadiaris 2018; Liu et al. 2017; Zhu et al. 2017a). They indeed deliver some performance gains; however, it is really a difficult task to obtain accurate human auxiliary information, such as detailed body parts and meticulous body poses, in the presence of those negative factors. Meanwhile, the traditional attention mechanism is prone to confusing areas, named distractions, which are caused by surrounding people, similar objects and puzzling backgrounds, as shown in Figure 2. These facts suggest space for improvement.

This paper proposes a novel method, namely Distraction-Aware HAR (Da-HAR), dealing with prediction of attributes of target persons in the wild (*i.e.* in the general domain). It emphasizes the importance of attribute-related region localization and enhances it by a coarse-to-fine attention mechanism, which largely reduces irrelevant distraction areas and substantially strengthens the following feature learning procedure. Specifically, at the coarse step, a self-mask block is designed to distill consensus information from extracted CNN features at different scales and highlight the most salient regions, in order to learn a rough distraction-aware mask from the input image. At the fine step, the traditional attention mechanism (Xu et al. 2015) is integrated by introducing a mask branch to further refine the distraction-aware information. This branch functions in a multi-task manner to boost results of classification and segmentation through making use of their interactions. In addition to distraction-aware attribute localization, considering that human attributes naturally exist at different semantic levels, we jointly use the features extracted from multiple layers of the deep network to improve the discriminative power. Comprehensive experiments are carried out on two major public benchmarks, *i.e.* WIDER-Attribute and RAP, and state-of-the-art results are reached, which clearly illustrate the competency of the proposed approach.

Related Work

Due to extensive real-world applications, HAR has drawn many attentions in recent years, with the scores on main benchmarks (Li et al. 2016b; Deng et al. 2014; Li et al. 2016a) continuously improved.

Early work focuses on extracting more discriminative features. (Cao et al. 2008) utilizes Histogram of Oriented Gradients (HoG) to represent person appearances and employs an ensemble classifier for gender recognition. (Joo, Wang, and Zhu 2013) builds a rich dictionary of detailed human parts based on HoG and color histograms, handling more attributes. CNN models, such as GoogLeNet (Szegedy et al. 2015), ResNet (He et al. 2016), and DenseNet (Huang et al. 2017), dominate this areas with stronger features and better scores.

Recently, increasing efforts have been made to improve attribute localization and thus boost the HAR performance. These methods can be classified into two categories. One explores auxiliary cues of other tasks, *e.g.* body part detection or annotation (Fabbri, Calderara, and Cucchiara 2017), pose estimation (Park, Nie, and Zhu 2017; Li et al. 2018), and human parsing (Tan et al. 2019), while the other introduces the attention mechanism to underline more important regions (Zhu et al. 2017a; Liu et al. 2017; Sarafianos, Xu, and Kakadiaris 2018; Tan et al. 2019).

(Fabbri, Calderara, and Cucchiara 2017) leverages body part annotation to calculate the pose-normalized feature maps of the head-shoulder, upper body, and lower body parts, respectively. The three feature maps are then employed in prediction and the optimal result is combined with that of the entire person to generate final decision. (Park, Nie, and Zhu 2017) jointly infers human poses and attributes in a sparse graph, which is built based on the annotated keypoints of the target person. (Li et al. 2018) transfers the knowledge learned from an off-the-shelf human pose estimation network and integrates pose information into attribute recognition to boost the performance. (Tan et al. 2019) simultaneously performs attribute recognition and human parsing in a multi-task learning manner, which further refines the features by parsing results. Unfortunately, target persons in HAR often have arbitrary variations in pose, occlusion, and background, *e.g.* in the surveillance and general domains. In this case, the clues of the auxiliary tasks mentioned above are no longer accurately available, making those methods unstable.

(Zhu et al. 2017a) firstly applies the attention mechanism to attribute recognition and discovers that the confidence-weighted scheme benefits most to recognition scores. (Sarafianos, Xu, and Kakadiaris 2018) follows this intuition and extends the regular attention scheme to a multi-scale attention-aggregated one. (Liu et al. 2017) learns multiple attentive masks corresponding to features at different layers and adds them back on the original features to form

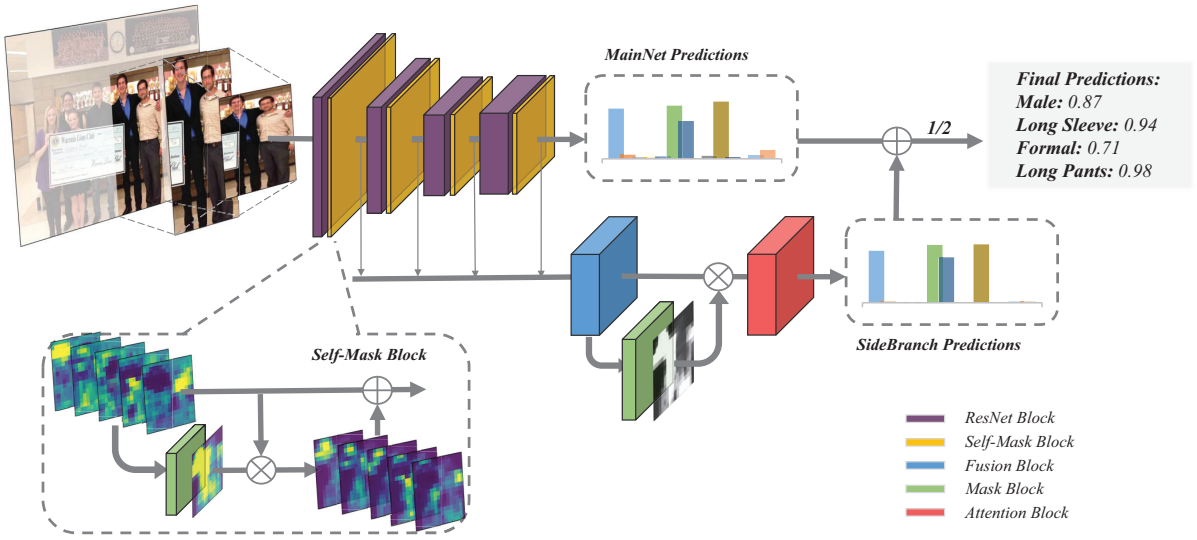

Figure 3: Overall framework of our Da-HAR. The Main Net (*purple*) follows the structure of ResNet-101. We insert several *self-mask blocks (yellow)* between different layers in the backbone network to learn the distraction awareness. Features from different levels are collected and fused by a *fusion block (blue)*, and the result is then sent to a side prediction branch, which contains a *mask block (green)* and an *attention block (red)* to guide the network to better concentrate on the attribute-related regions.

up attentive ones. (Tan et al. 2019) combines parsing attention, label attention, and spatial attention to ameliorate the accuracy of attribute localization. In particular, spatial attention learns several attention masks to better concentrate on attribute-related regions for prediction. However, these methods conduct attention learning on the entire input patch without any constraint, and the weights or masks generated are thus sensitive to confusion areas incurred by surrounding people, similar objects, and puzzling background, as shown in Fig. 2, making their results problematic.

The proposed method presents a novel approach, which tends to make use of both the advantages of the two groups of methods by inducing additional clues from person segmentation into attention learning in a progressive way. It localizes more accurate attribute regions by largely eliminating distraction areas through a coarse-to-fine attention scheme, and builds more powerful attribute features by integrating the ones extracted from multiple layers of the deep network.

Methodology

Overall Framework

The overall framework of our approach is illustrated in Figure 3. It consists of a main network as well as a side branch. Given an image patch with a target person, the proposed method processes it through both the main network and the side branch and their attribute predictions are combined to provide final results.

In the main network, we adopt ResNet-101 as the backbone, following recent studies (Zhu et al. 2017a; Sarafianos, Xu, and Kakadiaris 2018). We then add the self-mask block to the backbone to generate the features that are enhanced by the learned coarse distraction awareness. The features finally pass through a global average pooling layer and deliver Main Net predictions. In the side branch, we collect and combine the features at multiple layers from the backbone. These aggregated features are fed to the masked attention block, which further refines the previous distraction awareness. The same structure for classification is used to produce Side Branch predictions. At last, the Main Net and Side Branch predictions are averaged for decision making.

Coarse-to-fine Distraction-Aware Learning

As mentioned in Sec. 1, distractions with the patterns or appearances that are similar to the target may appear in the given image in the form of surrounding people, similar objects, and puzzling backgrounds (see Figure 2), which tend to induce false activations during the recognition process. The case even degenerates in real-world crowded scenes, such as assembly and campaign. Therefore, it is a difficult task to separate the target from distractions through bounding boxes. To deal with this issue, we propose a coarse-to-fine attention mechanism to progressively learn distraction awareness and substantially strengthen attribute features.

Coarse Distraction-Aware Learning. In the first stage, we roughly approximate the location of the target person by using the provided image-level labels. As we know, labels are assigned to the target person, to whom the highlighted features (activated neurons) are correlated. Therefore, a rough localization can be obtained by combing the features according to a weighted sum rule. To verify this intuition, we sum up all the channels of the output feature from the 4th layer or the 5th layer and employ the median value as

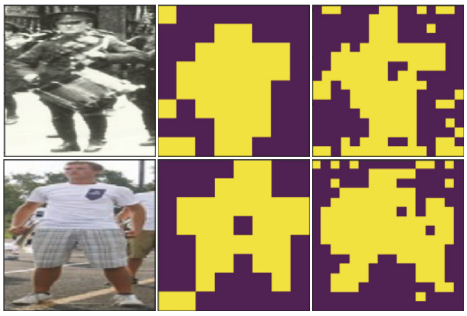

Figure 4: Visualization of summed features. By summing up the output features at the 4th layer (*3rd column*) or the 5th layer (*2nd column*) along the dimension of the channels and using the median value as a threshold for binarization, the approximate outline of the target person is rendered.

a threshold to binarize the summed feature map. As we can see in Figure 4, an approximate outline of the target person.

Based on the analysis above, we propose a self-mask block to learn the distraction-awareness. The structure of our self-mask block is shown in Figure 3. It contains a mask block to predict the saliency information, which is employed to optimize the input features to filter out the unconcerned regions. The mask block is constructed similar to a residual block, which is a stack of three 1×1 convolution layers and some batch normalization and activation layers. Since the self-mask blocks access the features at the corresponding layers, they only capture coarse distraction information.

Fine Distraction-Aware Learning. To further improve the localization accuracy of attribute-related regions, we propose a masked attention block. Since this block has access to the features collected from different layers, which contain the coarse distraction-awareness, it makes refinement to build more precise distraction-awareness.

A confidence-weighted attention mechanism (Zhu et al. 2017a) is exploited in the side branch to guide the network to explicitly focus on the attribute-related regions. However, recall that the essence of the attention mechanism is a re-weighting process on the give features, and the attention mechanism thus cannot rectify the false activations induced by distractions. To introduce distraction-aware learning to the regular attention mechanism, we design a masked attention block by integrating a mask branch after multi-level feature fusion (described in the next sub-section), as shown in Figure 3. This mask is employed to preprocess the fused feature before it is passed to the attention module. Note that the mask block constructed here is different from that in the self-mask block. We add two residual blocks on the top of the mask block so that it works in a multi-task manner, dealing with recognition and segmentation rather than simply distilling the consensus information.

Multi-Level Feature Aggregation

Features at multiple layers are collected and leveraged in the side branch to enhance the representation ability, as human attributes correspond to different semantic levels. To be specific, some attributes, such as long hair, wearing logo and stripes, are highly correlated to low-level features, *e.g.* colors, edges and local textons, which are encoded by the bottom layers of the network. Meanwhile, other attributes, like wearing T-shirt, sunglasses and jeans, are object-related, which are relevant to mid-level semantics extracted from deeper layers. In addition, a few attributes are at semantically high level and related to the features from top layers with several co-existing evidences. For example, gender is linked to body shapes and dressing styles, while dressing formally relates to ties and suits.

As features at different layers possess different spatial sizes and numbers of channels, a fusion block is proposed to balance the spatial resolutions and channel widths of these features. This block consists of three components: a 1×1 convolution layer for channel reduction to the channel widths; a re-sampling layer for scale adjustment to reconcile spatial resolutions; and a layer of stacked residual blocks to alleviate the aliasing effect of up-sampling.

Training Scheme

Loss Functions. The Binary Cross-Entropy (BCE) loss is adopted in model training, formulated as:

o

$$L_b(\hat{y}_p, y) = -\sum_{c=1}^{C} log(\sigma(\hat{y}_p^c))y^c + log(1-\sigma(\hat{y}_p^c))(1-y^c), \quad (1)$$

where C is the total number of attributes, $\hat{y}_p^c$ and y^c correspond to the predicted result and ground-truth label for each attribute c, respectively, and $\sigma(\cdot)$ represents the sigmoid activation function.

When the total number of concerned attributes increases, the influence of the class imbalance problem can no longer be neglected. We thus also employ the weighted BCE-loss (Yu et al. 2016) as:

$$\begin{aligned} L_w(\hat{y}_p, y) = -\sum_{c=1}^{C} \frac{1}{2\omega_c} log(\sigma(\hat{y}_p^c))y^c + \\ \frac{1}{2(1-\omega_c)} log(1-\sigma(\hat{y}_p^c))(1-y^c), \end{aligned} \quad (2)$$

where ω_c is the positive sample ratio of attribute c.

Supervision without Ground-Truth. If the ground-truth masks are given in the HAR datasets, our training can be easily performed. Unfortunately, the public HAR datasets rarely provide pixel-level annotations. Besides, pixel-level manual annotations are time-consuming. Therefore, we make use of the semantic segmentation results, which are obtained by an off-the-shelf segmentation technique, as our pixel-level ground truths, since it reports excellent performance.

Specifically, we train the segmentation network on the MS-COCO dataset, where FPN (Lin et al. 2017) is employed as the backbone network. The P2 feature is utilized as the final representation for segmentation. We crop target persons by bounding boxes and resize them to 224×224 as input. For training the masks, we only exploit the annotations of

Method	*Male*	*Long Hair*	*Sunglasses*	*Hat*	*T-shirt*	*Long Sleeve*	*Formal*	*Shorts*	*Jeans*	*Long Pants*	*Skirts*	*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	
R-CNN *ICCV'15*	94	81	60	91	76	94	78	89	68	96	80	72	87	55	80.0
R*CNN *ICCV'15*	94	82	62	91	76	95	79	89	68	96	80	73	87	56	80.5
DHC *ECCV'16*	94	82	64	92	78	95	80	90	69	96	81	76	88	55	81.3
VeSPA *BMVC'17*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.4
CAM *PRL'17*	95	85	71	94	78	96	81	89	75	96	81	73	88	60	82.9
ResNet-101 *CVPR'16*	94	85	69	91	80	96	83	91	78	95	82	74	89	65	83.7
SRN* *CVPR'17*	95	87	72	92	82	95	84	92	80	96	84	76	90	66	85.1
DIAA *ECCV'18*	96	88	74	93	83	96	85	**93**	**81**	96	85	78	90	68	86.4
Da-HAR **(Ours)**	**97**	**89**	**76**	**96**	**85**	**97**	**86**	92	**81**	**97**	**87**	**79**	**91**	**70**	**87.3**

Table 1: Quantitative comparison in terms of mAP between our proposed approach and eight counterparts on the WIDER-Attribute dataset. Note that the asterisk mark next to SRN indicates that the method is re-implemented by (Sarafianos, Xu, and Kakadiaris 2018), because the original work includes the validation set for training while the others do not.

the target person and discard all the unnecessary labels. According to our experiments, the generated segmentation results well serve as the ground truths for the masked attention block in Da-HAR, except for some cases when the target person is mostly occluded by surrounding people or objects.

Experiments

We evaluate our approach on two major public datasets, namely WIDER-Attribute and RAP. The WIDER-Attribute dataset consists of 13,789 images, where 57,524 instances are labeled with corresponding bounding boxes and 14 attribute categories. The RAP dataset is smaller than WIDER-Attribute, but it contains a larger number of attribute categories. It has a total number of 41,585 cropped images from 26 surveillance cameras, and the images are annotated with 69 attribute categories.

Training Details

We utilize a ResNet-101 model (He et al. 2016) pre-trained on the ImageNet dataset (Deng et al. 2009), as the backbone of our Da-HAR. All the proposed blocks, *i.e.*, the self-mask block, the feature fusion block, and the masked attention block, are initialized with Gaussian noise ($m = 0, \sigma = 0.01$).

To avoid over-fitting, the data augmentation strategies employed in (Zhu et al. 2017a) are adopted in our training. We firstly resize the input images to 256×256. Then, the resized images are cropped at four corners and the center, with the cropping width and height randomly chosen from the set $\{256, 224, 192, 168, 128\}$. At last, the cropped images are resized to 224×224. Note that we keep the ratio of height/width at 2 for the RAP dataset. Random horizontal flipping and color jittering are also applied for the data augmentation.

The stochastic gradient descent algorithm is utilized in the training process, with a batch size of 32, a momentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set to 0.003, and gamma is set to 0.1. Our model is trained on a single NVIDIA 1080Ti GPU. Two different strategies are used in inference. One directly resizes the input images to 256×256, while the other introduces an averaged result of five-crop evaluation.

Results

WIDER-Attribute. On this dataset, mean Average Precision (mAP) is employed as the metric. We compare our Da-HAR with the latest methods, including R-CNN (Girshick 2015), R*CNN (Gkioxari, Girshick, and Malik 2015), DHC (Li et al. 2016b), CAM (Guo, Fan, and Wang 2017), VeSPA (Sarfraz et al. 2017), SRN (Zhu et al. 2017a), DIAA (Sarafianos, Xu, and Kakadiaris 2018) and a fine-tuned ResNet-101 network (He et al. 2016). Note that SRN results are quoted from (Sarafianos, Xu, and Kakadiaris 2018), where the method is re-implemented without using the validation set for fair comparison.

Table 1 displays the results of different methods in details. We can see that our approach outperforms all the existing methods on the WIDER-Attribute dataset. Compared to DIAA, which reports the previous state-of-the-art score, Da-HAR achieves an mAP gain of 0.9%. Specifically, the results of some attributes related to accessories (*e.g.*, Hat and Sunglasses) and with strong local patterns (*e.g.*, T-shirt, Skirt and Plaid) increase more than 1%. Since these attributes usually have strong activations within small regions, they are easier to be influenced by surrounding distractions. By introducing the coarse-to-fine attention mechanism, Da-HAR proves effective in reducing such interferences and thus strengthening features. In particular, the proposed method shows the largest improvement (3%) on the Hat attribute, even though the baseline is relatively high (93%). In general, Hat is a frequently appeared attribute and its prediction seriously suffers from distractions. It highlights the advantage of the proposed method.

RAP. On this dataset, we follow the standard protocol (Li et al. 2016a) and only report the prediction results on 51 attributes whose positive sample ratios are higher than

Method	*loss*	*mA (%)*	*Acc (%)*	*Prec (%)*	*Rec (%)*	*F1 (%)*
HPNet *ICCV'17*	*s.*	76.12	65.39	77.53	78.79	78.05
JRL *ICCV'17*	*o.*	77.81	-	78.11	78.98	78.58
VeSPA *BMVC'17*	*s.*	77.70	67.35	79.51	79.67	79.59
WPAL *BMVC'17*	*w.*	81.25	50.30	57.17	78.39	66.12
GAM *AVSS'17*	*o.*	79.73	**83.79**	76.96	78.72	77.83
GRL *IJCAI'18*	*w.*	81.20	-	77.70	80.90	79.29
LGNet *BMVC'18*	*s.*	78.68	68.00	80.36	79.82	80.09
PGDM *ICME'18*	*w.*	74.31	64.57	78.86	75.90	77.35
VSGR *AAAI'19*	*o.*	77.91	70.04	82.05	80.64	**81.34**
RCRA *AAAI'19*	*w.*	78.47	-	82.67	76.65	79.54
IA^{2}Net *PRL'19*	*f.*	77.44	67.75	79.01	77.45	78.03
JLPLS *TIP'19*	*o.*	81.25	67.91	78.56	81.45	79.98
CoCNN *IJCAI'19*	*o.*	81.42	68.37	81.04	80.27	80.65
DCL *ICCV'19*	*m.*	83.70	-	-	-	-
Da-HAR (**Ours**)	*o.*	73.78	68.67	**84.54**	76.84	80.50
	w.	**84.28**	59.84	66.50	**84.13**	74.28
	m.	79.44	68.86	80.14	81.30	80.72

Table 2: Quantitative results of our proposed method and fourteen counterparts on the RAP dataset. *s.*, *o.*, *w.*, *f.*, and *m.* represent the Softmax loss, Original BCE-loss, Weighted BCE-loss, Focal BCE-loss, and a Mixed loss, respectively.

1%. Multiple metrics are used, involving mean Accuracy (mA), instance Accuracy (Acc), instance Precision (Prec), instance Recall (Rec), and instance F1 score (F1). Refer to (Li et al. 2016a) for their definitions and explanations. We compare our approach with 14 existing counterparts, including HPNet (Liu et al. 2017), JRL (Wang et al. 2017), VeSPA (Sarfraz et al. 2017), WPAL (Yu et al. 2016), GAM (Fabbri, Calderara, and Cucchiara 2017), GRL (Zhao et al. 2018), LGNet (Liu et al. 2018), PGDM (Li et al. 2018), VSGR (Li et al. 2019), RCRA (Zhao et al. 2019), I^{2}ANet (Ji et al. 2019), JLPLS (Tan et al. 2019), CoCNN (Han et al. 2019), and DCL (Wang et al. 2019), as shown in Table 2. The samples in the RAP dataset are collected from real world surveillance scenarios, and compared to the ones in WIDER-Attribute, there are less distractions. Under this circumstance, our method still reaches very promising results compared to the state-of-the-art methods.

Since the problem of class imbalance in RAP is very severe, the performance fluctuates under different loss functions. Da-HAR effectively filters out distractions and thus reduces false positives, showing the superiority in Precision with the highest score of 84.54% under BCE-loss. As the original BCE-loss emphasizes the majority attributes, the trained network tends to induce strong biases on the minority ones, which impairs Recall. The weighted BCE-loss assigns penalties to the minority attributes and enforces the network to better distinguish them. In this case, the highest instance Recall of 84.13% is achieved , along with the highest mean Accuracy of 84.28%. However, the discriminative information in the majority attributes is suppressed, which leads to a Precision drop. At last, a mixed loss function makes a better trade-off between Precision and Recall and delivers a moderate F1 score of 80.72%.

Ablation Study

To validate our contributions, we further perform ablation studies on the WIDER-Attribute dataset. The images are acquired in unconstrained scenarios, including marching, parading, cheering and different ceremonies, where the target person is more likely to appear in crowds, *i.e.*, distractions occur more frequently in cropped input images. Therefore, this dataset is suitable for analyzing the proposed method.

Multi-Level Feature	Self-Mask	Masked Attention	Ignore 0-labeled	**mAP (%)** Crops / Full
				85.3 / 86.2
✓				85.7 / 86.5
	✓			85.8 / 86.6
		✓		86.0 / 86.8
✓	✓	✓		86.2 / 87.1
✓	✓	✓	✓	86.5 / **87.3**
✓	✓	✓	✓	**87.2 / 88.0**

Table 3: Ablation Study on the WIDER-Attribute dataset. We re-implement SRN (Zhu et al. 2017a) without its spatial regularization module as our baseline, whose results are reported in the first row. The last row displays the results obtained by adding the validation subset to the training process.

Sub-Module Analysis. Here, we investigate the contribution of each module to the final results of Da-HAR in Table 3. We re-implement SRN (Zhu et al. 2017a) without its spatial regularization module to serve as our baseline network, which achieves an mAP of 85.3% (obtained by using the average results of five-crop evaluation). If we utilize the uncropped (full) patch and resize it to 256×256 as the input, the baseline increases to 86.2%, as the uncropped patch contains more information. Our feature fusion block boosts the performance to 85.7%/86.5%, which demonstrates the effectiveness of leveraging multi-level semantic clues. When the self-mask block is applied only, an mAP of 85.8%/86.6% is reached by learning the coarse distraction-awareness. The masked attention block, *i.e.*, the fine distraction-aware learning step, further improves the mAP to 86.0%/86.8%. These facts underline the necessity of the coarse-to-fine attention scheme. By integrating all the three components, the mAP rises to 86.2%/87.1%. We also observe that the mAP score grows to 86.5%/87.3%, if we ignore 0-label attributes at the training phase, which are previously treated as negatives in the literature. This step makes data more consistent in training, in spite of losing some samples. If the validation set is added to the training one as (Zhu et al. 2017a) does, an mAP of 87.2%/88.0% is obtained.

Masked Attention Analysis. To further verify the effectiveness of the coarse-to-fine distraction-aware feature learning and demonstrate the impact of the auxiliary supervision from segmentation, we visualize the saliency mask, *i.e.*, the output of the masked attention module in the side branch in Figure 5. The visualized masks are trained with and without the auxiliary supervision. For convenience, a cropped input image and the corresponding masks trained without/with supervision are gathered to form a triplet of images. These triplets are arranged into three panels, which are separated by two dashed lines, according to their different levels of

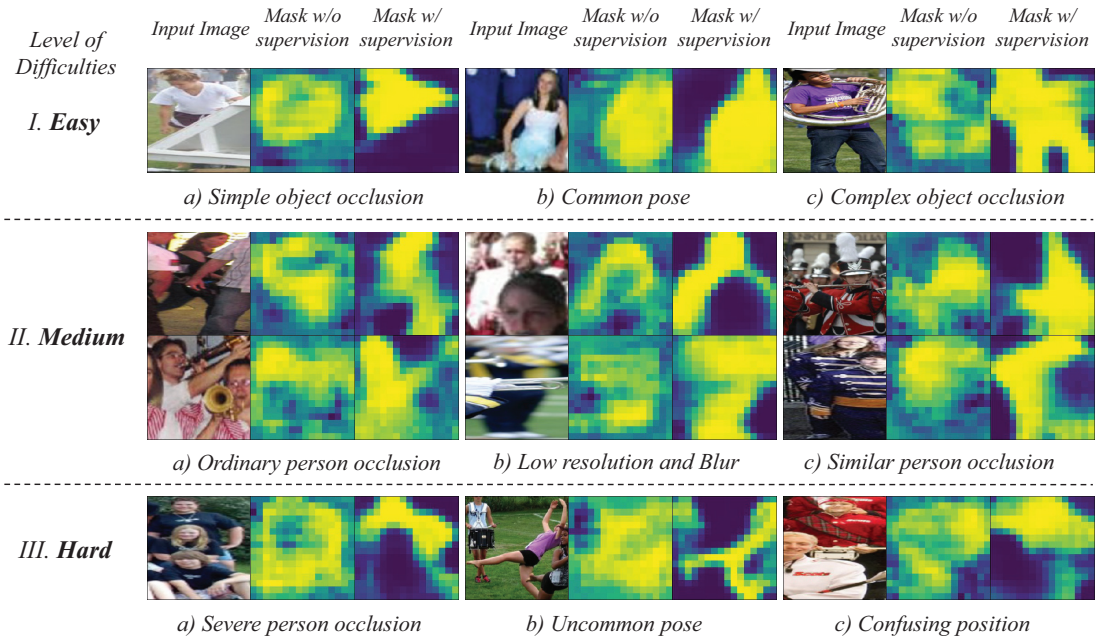

Figure 5: Visualization of Attentive Mask. The visualized masks are trained with and without the auxiliary supervision.

difficulties. In each panel, they are grouped into three zones, with respect to the challenges.

The first panel consists of images with simple object occlusion (*I-a*), common pose (*I-b*), and moderate occlusion by complex objects (*I-c*). In this case, the saliency masks generated without supervision are well estimated with slight noises. When we use proper supervision, the generated masks are obviously better.

The second panel contains images with severe occlusions by ordinary people (*II-a*), low resolution and blur (*II-b*), and occlusions by people having similar appearances (*II-c*). Although the distractions are more severe compared to those in the first panel, our mask still locates the target person without supervision. If supervision is included in training, our method achieves a very accurate localization.

The images in the third panel are considered to be hard for HAR with heavy occlusions by people having extremely similar appearances (*III-a*), irregular pose (*III-b*), and severe occlusions with confusing positioning (*III-c*). The white jacket (*III-c*) are wrongly identified as the pants of the target person, who is dressed in red. Under such circumstances, the masks trained without supervision are not good, but the supervised ones are decent, which demonstrates the effectiveness of the pre-trained segmentation network.

Computational Complexity Analysis. To better understand our method, we calculate the overall network parameters and record the inference time. The baseline network has 45.4M parameters, and the inference time of a single image by a NVIDIA 1080Ti is 22.28ms. When all the blocks are added, the final network (Da-HAR) has 52.4M parameters in total and the inference time only slightly increases to 24.04ms.

Conclusion

In this paper, we introduce a distraction-aware learning method for HAR, namely Da-HAR. It underlines the necessity of attribute-related region localization and a coarse-to-fine attention scheme is proposed for this issue. A self-mask block is presented to roughly locate distractions at each layer, and a masked attention mechanism is designed to refine the coarse distraction awareness to further filter out false activations. We also integrate the multi-level semantics of human attributes to improve the performance. Extensive experiments are carried out on the WIDER-Attribute and RAP datasets and state of the art results are reached, which demonstrate the effectiveness of the proposed Da-HAR.

Acknowledgment

This work was partly supported by the National Key Research and Development Plan (No. 2016YFB1001002), the National Natural Science Foundation of China (No. 61802391), and the Fundamental Research Funds for the Central Universities.

References

Al-Halah, Z.; Stiefelhagen, R.; and Grauman, K. 2017. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 388–397.

Cao, L.; Dikmen, M.; Fu, Y.; and Huang, T. S. 2008. Gender recognition from body. In *MM*, 725–728.

Chen, Q.; Huang, J.; Feris, R.; Brown, L. M.; Dong, J.; and Yan, S. 2015. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 5315–5324.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *MM*, 789–792.

Fabbri, M.; Calderara, S.; and Cucchiara, R. 2017. Generative adversarial models for people attribute recognition in surveillance. In *AVSS*, 1–6.

Feris, R.; Bobbitt, R.; Brown, L.; and Pankanti, S. 2014. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 153–160.

Gao, L.; Huang, D.; Guo, Y.; and Wang, Y. 2019. Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention. In *MM*, 1340–1348.

Girshick, R. 2015. Fast r-cnn. In *ICCV*, 1440–1448.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with r* cnn. In *ICCV*, 1080–1088.

Guo, H.; Fan, X.; and Wang, S. 2017. Human attribute recognition by refining attention heat map. *PRL* 94:38–45.

Han, K.; Guo, J.; Zhang, C.; and Zhu, M. 2018. Attribute-aware attention model for fine-grained representation learning. In *MM*, 2040–2048.

Han, K.; Wang, Y.; Shu, H.; Liu, C.; Xu, C.; and Xu, C. 2019. Attribute aware pooling for pedestrian attribute recognition. *arXiv preprint arXiv:1907.11837*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Ji, Z.; He, E.; Wang, H.; and Yang, A. 2019. Image-attribute reciprocally guided attention network for pedestrian attribute recognition. *PRL* 120:89–95.

Joo, J.; Wang, S.; and Zhu, S.-C. 2013. Human attribute recognition by rich appearance dictionary. In *ICCV*, 721–728.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016a. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*.

Li, Y.; Huang, C.; Loy, C. C.; and Tang, X. 2016b. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 684–700.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2018. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *ICME*, 1–6.

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019. Visual-semantic graph reasoning for pedestrian attribute recognition. In *AAAI*, volume 33, 8634–8641.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Ling, H.; Wang, Z.; Li, P.; Shi, Y.; Chen, J.; and Zou, F. 2019. Improving person re-identification by multi-task learning. *Neurocomputing* 347:109–118.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 1096–1104.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 350–359.

Liu, P.; Liu, X.; Yan, J.; and Shao, J. 2018. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*.

Park, S.; Nie, B. X.; and Zhu, S.-C. 2017. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE TPAMI* 40(7):1555–1569.

Sarafianos, N.; Vrigkas, M.; and Kakadiaris, I. A. 2017. Adaptive svm+: Learning with privileged information for domain adaptation. In *ICCV*, 2637–2644.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2018. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, 680–697.

Sarfraz, M. S.; Schumann, A.; Wang, Y.; and Stiefelhagen, R. 2017. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*, 1–9.

Tan, Z.; Yang, Y.; Wan, J.; Wan, H.; Guo, G.; and Li, S. Z. 2019. Attention-based pedestrian attribute analysis. *IEEE TIP* 28(12):6126–6140.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 5079–5087.

Wang, X.; Zhang, T.; Tretter, D. R.; and Lin, Q. 2013. Personal clothing retrieval on photo collections by color and attributes. *IEEE TMM* 15(8):2035–2045.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, 531–540.

Wang, Y.; Gan, W.; Wu, W.; and Yan, J. 2019. Dynamic curriculum learning for imbalanced data classification. *arXiv preprint arXiv:1901.06783*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yu, K.; Leng, B.; Zhang, Z.; Li, D.; and Huang, K. 2016. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*.

Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; and Jin, X. 2018. Grouping attribute recognition for pedestrian with joint recurrent learning. In *IJCAI*, 3177–3183.

Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; and Yan, C. 2019. Recurrent attention model for pedestrian attribute recognition. In *AAAI*, 9275–9282.

Zhu, F.; Li, H.; Ouyang, W.; Yu, N.; and Wang, X. 2017a. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, 5513–5522.

Zhu, J.; Liao, S.; Lei, Z.; and Li, S. Z. 2017b. Multi-label convolutional neural network based pedestrian attribute classification. *IVC* 58:224–229.