Localization Guided Learning for Pedestrian Attribute Recognition

Abstract

Pedestrian attribute recognition has attracted many attentions due to its wide applications in scene understanding and person analysis from surveillance videos. Existing methods try to use additional pose, part or viewpoint information to complement the global feature representation for attribute classification. However, these methods face difficulties in localizing the areas corresponding to different attributes. To address this problem, we propose a novel Localization Guided Network which assigns attribute-specific weights to local features based on the affinity between proposals pre-extracted proposals and attribute locations. The advantage of our model is that our local features are learned automatically for each attribute and emphasized by the interaction with global features. We demonstrate the effectiveness of our Localization Guided Network on two pedestrian attribute benchmarks (PA-100K and RAP). Our result surpasses the previous state-of-the-art in all five metrics on both datasets.

1 Introduction

Pedestrian attribute recognition is a high-demanding problem due to its wide applications in person re-identification, person retrieval, and social behavioral analysis. Boosted by the increasing demand, three pedestrian attribute benchmarks [7, 17, 20] are released to facilitate the revolution of pedestrian attribute recognition approaches. Inspired by the increasing amount of annotated data, a sequence of deep learning based methods [16, 20, 25, 28, 29] are proposed with remarkable advantages over previous SVM-based approaches. Most recently-proposed methods formulate pedestrian attribute recognition as a multi-label classification problem, and address it by optimizing a parameter-sharing convolutional network to exploit semantic relations among attributes in an adaptive manner, which achieves a huge success.

The performances of these deep learning methods are constrained by several factors. Firstly, the resolution of pedestrian images is not admirable. Specifically, the lowest resolutions of images are $36 \nrightarrow 92$ for RAP and $50 \nrightarrow 100$ for PA-100K. Moreover, the scales of

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

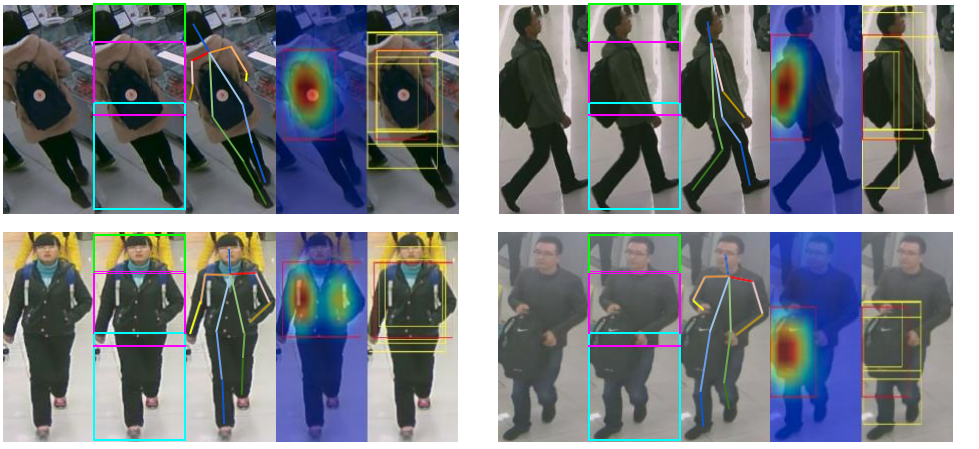

Figure 1: Pedestrian attribute recognition with local features relies on the accurate localization of attributes. Take backpack as an example, we present schematic diagrams of different methods, i.e. part-based method (b), keypoint-based method (c), activation map-based method (d) and our localization guided method (e, represented by top-5 proposals). Result suggests that the object is better localized in our method regardless of the variance in the viewpoints and the relative positions of objects. Meanwhile, our method is in favor of the object and its surrounding areas, compared with only the object in the class activation map.

attributes vary from the whole image (e.g. "fat", "thin") to small regions (e.g. "glasses", "shoes"). It could be especially challenging to classify some small attributes due to the information loss caused by down-sampling. Secondly, due to the divergence of the angles and distances between the pedestrians and the surveillance cameras, as well as the imperfection of pedestrian detection algorithms, the pedestrian images have great variation in shapes and viewpoints. The recognition difficulty increases due to the variation in viewpoints, as introduced by Sarfraz *et al*. [25]. For example, Figure 1(a) shows some examples of backpack recognition. The original images in column (a) show images from three viewpoints, and the backpack appears in hand in the last image. The large variance of relative positions between the pedestrians and these backpacks suggests that it is hard to localize objects only by its relative position regardless of the input image.

One intuitive solution to address the problems mentioned above is to apply local feature mining in order to fully explore the information from specific regions. Most existing CNN-based methods use either human part or pose information to extract local features for multiple pedestrian recognition tasks. For part information, one typical solution as introduced in [9, 18] is to split the image vertically into upper, middle and lower part, and feed each part into a distinct network for feature extraction. The features are merged by concatenating with a global feature vector learned from the whole image for end-to-end optimization. One problem with these methods is that the splitting of the body part is defined by a fixed, human-specific strategy that can be imprecise. It may lead to truncated parts for some images. To better exploit the human part in the images, pose information is introduced in [23, 35] that utilize a landmark model for body part localization. While due to the imperfection of human

detector quality and pose variations, the missing body part in the image could confuse such models. Another solution is to design a network based on attention models. Liu *et al*. [20] introduced a soft attention model that extract and concatenate multi-level global and local feature patterns for attribute classification.

The common constraint of the methods mentioned above is that they do not specifically locate attributes in the images, but use human-defined local feature extraction strategies. These methods surpassed ordinary results that only extracts global information, indicating that the exploited local information can boost the recognition of attributes to some extent. To overcome the drawbacks of previous methods and fulfill our expectations, we propose a novel network architecture, referred as Localization Guided Network (LG-Net) as shown in Figure 2 to fully exploit the attribute-specific local features for attribute predictions. Specifically, we modify the class activation map [36] to dynamically generate localization result for the accurate localization of attributes. The localization of a certain attribute is further projected to the local feature maps to achieve attribute-wise local feature extraction. The extracted local features are merged with global features in an adaptive manner. Compared with previous methods, our architecture could accurately address the locations of certain attributes, especially object-type attributes, and match the surrounding local features for location-specific attribute recognition. We evaluate our result on the two largest pedestrian datasets by five matrices, showing that our goal is achieved with a significant improvement in attribute prediction result.

2 Related Work

Pedestrian Attribute Recognition. Pedestrian attribute recognition is a multi-label classification problem that has been widely applied to the person retrieval and person re-identification. Earlier works address this problem with some traditional methods. Gray *et al*. [11] apply AdaBoost algorithm to address this problem. Prosser *et al*. [24], Layne *et al*. [14, 15], Deng *et al*. [7] and Li *et al*. [17] solve this problem by optimizing Support Vector Machines (SVMs) to classify single attributes. The recognition accuracy is significantly improved with the recent development of Deep Learning. Sudowe *et al*. [28] first introduced a convolutional neural network with parameter sharing in most layers to adaptively explore the semantic relations among attributes. The network is trained end-to-end with independent loss layers for each attribute.

Some recent methods like Yu *et al*. [32] and Liu *et al*. [20] capture multi-level features from the network to exploit global and local contents. Li *et al*. [16] design a weighted loss to address the problem of imbalanced positive and negative samples. Sarfraz *et al*. [25] introduce a model with view guidance to make view-specific attribute predictions so as to overcome the variance of patterns in different angles. Wang *et al*. [29] proposed a recurrent neural network to explore intra-person, inter-person, and inter-attribute context information to better exploit semantic relations among attributes.

Other methods further improve the model design with the pose or part information. Zhang *et al*. [34] explore poselets for part localization and attribute classification. Fabbri *et al*. [9], Li *et al*. [18] and Zhu *et al*. [37] split input image into fixed parts by some handcraft standard to extract features from certain part. Park *et al*. [23] and Zhao *et al*. [35] explore key-points of the person to dynamically capture the body pose for part split and local feature extraction. Yang *et al*. [31] use key-points to generate adaptive parts. These methods utilize different information to extract local features. But the attribute-specific localization infor-

mation is missing in these solutions, which makes these methods less robust to the variance of the relative position of the objects of interest.

Weakly Supervised Localization. Weakly supervised localization using only image-level supervision is an important problem to reduce the cost of annotations. Some methods [4, 6, 8, 13, 22, 27, 30] formulate this problem as multiple instance learning (MIL). Oquab *et al*. [21] utilize mid-level representations that solve object localization with CNN outputs. Oquab *et al*. [22] further develop a CNN that uses global max pooling on top of features to localize a point on objects. Similarly, Zhou *et al*. [36] introduce global average pooling to obtain a class activation map. Bounding boxes covering the highlight areas are adopted as the localization result. Another solution is to cluster the similar patterns among regions of interests. Song *et al*. [27] borrow the idea from nearest neighbor to select a set of adjacent windows with graph-based algorithms. Bilen *et al*. [3] propose a network containing two parallel branches with different normalization to perform classification and weakly supervised detection respectively. Bency *et al*. [2] compare different candidate localization to confirm boxes as localization results progressively. Bazzani *et al*. [1] and Singh *et al*. [26] enhance the performance by randomly hide some spatial regions for the classification network to capture more relevant patterns. Li *et al*. [19] extend this idea by filtering the highlight areas instead of random regions to perform accurate weakly-supervised segmentation.

3 Localization Guided Network Architecture

In order to extract and utilize the local features around the locations of attributes, we design a novel architecture for localization guided feature extraction and attribute classification, as shown in Figure 2. The network is composed of two separate branches, i.e. global branch and local branch. The global branch takes image-level input to generate localization for all attributes, and the local branch utilizes localization to predict the attributes. Both global and local branches are adapted from the *Inception-v2* [12] architecture for its great extensibility in multi-scale feature extraction.

3.1 Global Feature Extraction

The global branch takes the whole image as input and extracts class activation maps [36], which can provide guidance on localizing attributes.

Class Activation Box. Class activation box is obtained from the class activation map, acting as the locations of attributes in our method, as shown in Figure 2. Specifically, the original CAM method takes the weighted sum of the output feature maps as the class activation map of a class, where the weights are corresponding to the weights between classes and the output units of the global average pooling (GAP) layer. In our method, we implement this in an equivalent way, to enable calculating class activation maps in a forward pass during training and testing. We extract the weight from a pretrained classification model and use the extracted weight to initialize the class activation generator, implemented as a fixed 1×1 convolution layer on top of the global branch. The convolution layer dynamically produces c class activation maps for each image, where c is the number of attributes in our dataset. In order to keep the activation map unchanged, we fix the parameter updating in the global branch. After the class activation maps are obtained, an activation box is captured for each attribute by cropping the high-response areas of the corresponding activation map. The activation boxes act as the localization result in our method.

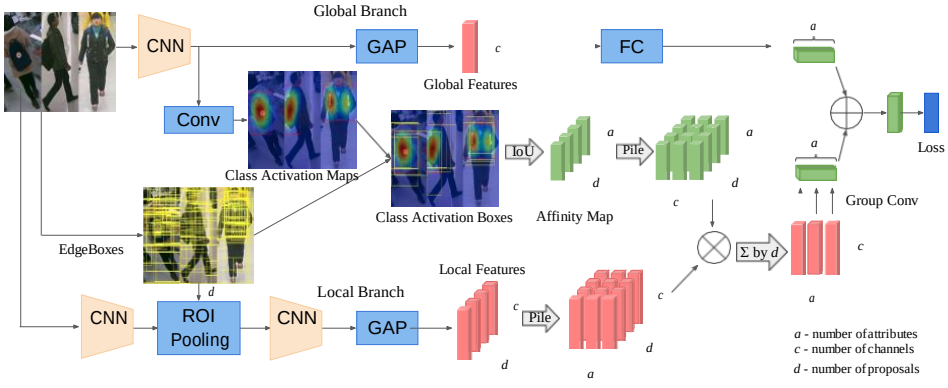

Figure 2: Our Localization Guided Network is composed of two branches, i.e. global branch and local branch. The global branch is a parameter-fixed branch, taking the whole image as input to generate attribute locations based on the class activation map. The convolution layer is fixed and initialized by the weights of the fully-connected layer in pretrained model. Localization results are used for calculating the weights to the local features for attribute-wise local feature selections. Weighted local features are projected to a single vector by attribute. The element-wise summation of global and local feature vectors are used for attribute predictions.

3.2 Local Feature Extraction

The local branch serves the purpose of extracting local features from the input image. In our method, we use EdgeBoxes [38] to generate region proposals. Specifically, for each input image $\mathbf{x}$, EdgeBoxes output a list of candidate regions $R = (R_1, ..., R_n)$ where R_i is a 5-dimensional array, containing a bounding box and a score for each box. We extract a fixed number of proposals for each image. The local branch extracts local features by an ROI pooling layer between *inception-5a* and *inception-5b*, such that a sufficient amount of information is provided for proposal processing while the overall computational cost is reasonable.

The extracted local features are fed into *inception-5b* and a following global average pooling for further processing. The outputs of the network are n feature vectors for each image where n is the number of proposals. All proposals are even in the feature extraction part. The n proposal regions could cover most possible locations of attributes, local features are extracted into separate vectors for future utilization.

3.3 Localization Guidance Module

In Section 3.2, we have extracted local features of all proposals. However, it is natural to consider that the proposals should not contribute equally, and different attributes should focus on different local features. For example, for the attribute 'wear glasses', the local features around head should contribute more, while for the attribute 'color of shoes', usually it relates to local features at the bottom. So the intuition is that we should decide how much each local feature should contribute to the classification of a certain attribute. Fortunately, in Section 3.1, we have obtained the class activation map for each attribute, which could serve

as a guide to determine the importance of the local features to different attributes.

So as to generate attribute-specific feature vectors for attribute recognition, localization guidance is expected beyond local feature extraction. We design a Localization Guidance Module that takes input information from both global and local branches to map local feature vectors with attributes by the spatial affinity between the proposals and the class activation boxes. In the following part, we will introduce the localization of attributes, the affinity calculation between ROIs and locations of attributes, together with the fusion strategy so as to generate the image-level attribute-specific information with local feature maps.

CAM-ROI Affinity Map. Different proposals are of variable importance to different attributes. To address the important regions for each attribute, class activation boxes generated by CAMs provide the information about the importance of proposals. In our method, the affinity between class activation boxes and proposals are calculated as the IoU (Intersection over Union) between the two bounding boxes as shown in Figure 2, due to the need of eliminating the influence of proposals with a little intersection with the object locations and punish the proposals that are too large to focus on a specific region of the attribute.

Formally, denoting c the number of classes in a dataset and $d = |R|$ the number of proposals, for each given image, two shortlists of bounding boxes are associated with length c generated from CAM [36] and d generated by EdgeBoxes [38] respectively. In order to capture the proposal boxes with high correlations with attribute localization, a CAM-ROI Affinity Map is expected. We formulate the Affinity as follows:

$$A_{i,j} = \frac{C_i \cap D_j}{C_i \cup D_j}$$

where C_i is the i-th element of the CAM boxes and D_j is the j-th element of the ROI regions. The affinity between two proposals is calculated by the Intersection over Union (IoU).

An affinity map in $c \times d$ is calculated and linearly normalized to weight the local feature vectors for further predictions. There is no need for the affinity function to be differentiable since the class activation maps are generated from a fixed branch.

Localization Guidance. The shapes of weighted local feature tensors are determined by the numbers of proposals and the number of attributes in consideration. While the shape of the global feature is dependent on the number of convolution kernels in the last layer. So as to take the element-wise summation from both global and local branch, a fusion module is designed to combine the output of two branches, while keeping the correlations between the proposals and the attributes consistent.

The Localization Guided Network is composed of a convolutional neural network with ROI pooling, followed by a localization guiding block. The localization guiding block is introduced as follows. For each sample, the input is an image, denoted as I, and its associated proposals R. The global stream would produce an affinity map:

$$A = f_g(I; q_g),\ A \in \mathbb{R}^{d \times c \times 1}$$

Meanwhile, the feature vector obtained from the local stream after global pooling is reshaped into:

$$X = f_l(I, R, q_l),\ X \in \mathbb{R}^{d \times 1 \times k}$$

Where f_g, f_l are the two global and local streams of the LG-Net and q the corresponding parameters. So as to apply element-wise weighting, A and X are both tiled into $\mathbb{R}^{d \times c \times k}$ and produced as the localization guided local feature matrices $\hat{X} \in \mathbb{R}^{d \times c \times k}$.

The local features need to be merged for image-level recognition. In our method, a non-weighted summation of all proposals is used as the image-level local feature tensor considering the scale. The image-level feature tensor preserves the attribute-wise information in the format of a 1024 dimensional feature vector for each attribute of an image. To obtain a c dimensional local feature vector for each image, each 1024 dimensional feature vector is mapped to a single value by an adaptively weighted sum. The local features are then mapped to an attribute-wise local feature vector $\hat{Y}_l \in \mathbb{R}^c$.

For the fusion of global and local features, we apply element-wise sum to the two mentioned feature matrices:

$$\hat{Y} = \hat{Y}_l \bigoplus \hat{Y}_g$$

The summation act as the classification score for each attribute supervised by a weighted sigmoid cross entropy loss, which takes image-level ground truth to update parameters by weighted gradient back-propagation. The gradient values of the positive samples are multiplied by a weighting factor w_c to overcome the data imbalance as introduced in [16].

3.4 Implementation Details

In this section, we use a to denote the number of attributes of interest, and c to denote the number of feature maps in the last layer of the fundamental convolutional network to maintain consistency with Figure 2. The LG-Net is trained in two stages. In the first training stage, an Inception-v2 classification network is trained from the pedestrian attribute benchmarks, initialized by an ImageNet pretrained model. The weight outcome in the first stage is denoted as q_a. The weight of the fully-connected layer in q_a is extracted as $q_a(FC) \in \mathbb{R}^{a*c}$.

In the second training stage, q_a is acting as the initialization of the global branch of a LG-Net, as shown in Figure 2. A convolution layer of size 1×1, stride 1 with c units is initialized by $q_a(FC)$ on top of the feature maps of the global branch. This layer generates class activation maps for each image in the forward pass of the network. The network parts mentioned above in the second stage are fixed without parameter updating. This setting is to ensure the accurate production of class activation boxes for the ultimate localization result.

The LG-network is optimized by Stochastic Gradient Descent (SGD) [5] with an initial learning rate of 0.02 and a weight decay of 0.005. The learning rate decays by 0.1 every 20 epochs. We train the network for 50 epochs and select the best model based on the performance on the validation set. For each image, we select top 100 proposals generated by EdgeBoxes [38] and post-processed by Non-Maximum Suppression. The activated region for an attribute is defined as the pixels whose activation value on the activation map is greater than 0.2 times the maximum activation value in the image for the particular attribute.

4 Experiment

4.1 Benchmark Data

RAP [17] and PA-100K [20] are the two largest public pedestrian attribute datasets. The Richly Annotated Pedestrian (RAP) dataset contains 41,585 images collected from indoor surveillance cameras. Each image is annotated with 72 attributes, while only 51 binary attributes with the positive ratio above 1% are selected for evaluation. There are 33,268 images for the training set and 8,317 for testing. PA-100K is a recent-proposed large pedestrian attribute dataset, with 100,000 images in total collected from outdoor surveillance cameras. It

Dataset	RAP					PA-100K				
Method	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
ELF+SVM [24]	69.94	29.29	32.84	71.18	44.95	-	-	-	-	-
CNN+SVM [17]	72.28	31.72	35.75	71.78	47.73	-	-	-	-	-
ACN [28]	69.66	62.61	80.12	72.26	75.98	-	-	-	-	-
DeepMar [16]	73.79	62.02	74.92	76.21	75.56	72.70	70.39	82.24	80.42	81.32
HP-Net [20]	76.12	65.39	77.33	78.79	78.05	74.21	72.19	82.97	82.09	82.53
JRL [29]	77.81	-	78.11	78.98	78.58	-	-	-	-	-
VeSPA [25]	77.70	67.35	79.51	79.67	79.59	76.32	73.00	84.99	81.49	83.20
Inception-v2 [12]	75.43	65.94	79.78	77.05	78.39	72.65	71.56	84.12	80.30	82.17
LG-Net	**78.68**	**68.00**	**80.36**	**79.82**	**80.09**	**76.96**	**75.55**	**86.99**	**83.17**	**85.04**

Table 1: Quantitative Comaprison with State-of-the-Art. Our method surpasses previous state-of-the-art by all five metrics on both datasets.

is split into 80,000 images for the training set, and 10,000 for the validation set and 10,000 for the test set. This dataset is labeled by 26 binary attributes. The common features existing in both selected dataset is that the images are blurry due to the relatively low resolution and the positive ratio of each binary attribute is low.

Following the settings of previous arts, we apply both *label-based* evaluation by calculating the mean Accuracy (mA) by averaging the accuracy on the positive and negative samples, and *sample-based* metrics including accuracy, precision, recall, and F1. These metrics are applied to our comparison on both datasets.

4.2 Quantitative Comparison with Prior Arts

We demonstrate the effectiveness on both datasets by keeping the same settings with previous methods. Table 1 shows the comparison of our approach with all available public datasets. Our proposed method is named in LG-Net, referring to the Localization Guided Network. We also list the result of our Inception-v2 baseline model.

The result suggests that our LG-Net surpasses all previous methods in both label-based and sample-based metrics on RAP dataset. Since PA-100K is a newly proposed dataset, it is less likely to compare all the previous pedestrian attribute methods. Following the evaluation method in [20], we replicate the state-of-the-art method, known as VeSPA on PA-100K, and compare our result with previous state-of-the-art and previously released results. Our LG-Net significantly surpasses VeSPA and earlier methods, suggesting that our method achieves the state-of-the-art result on PA-100K.

4.3 Qualitative Evaluation

To highlight the performance of the attribute-specific localization result of the proposed method, we compare the mean accuracy of 51 attributes in RAP dataset between our LG-Net and VeSPA, the previous state-of-the-art, as shown in Figure 3. The result shows that our method performs better than VeSPA in most attributes, especially attached objects, for example, "backpacks", "plastic bags" and "calling". The attached objects are closely related to our proposed localization guided method since attached objects are more sensitive to the relative positions and viewpoints. Moreover, our method performs slightly better in action recognition including "talking" and "gathering" because the LG-Net utilizes both global and local information for attribute predictions.

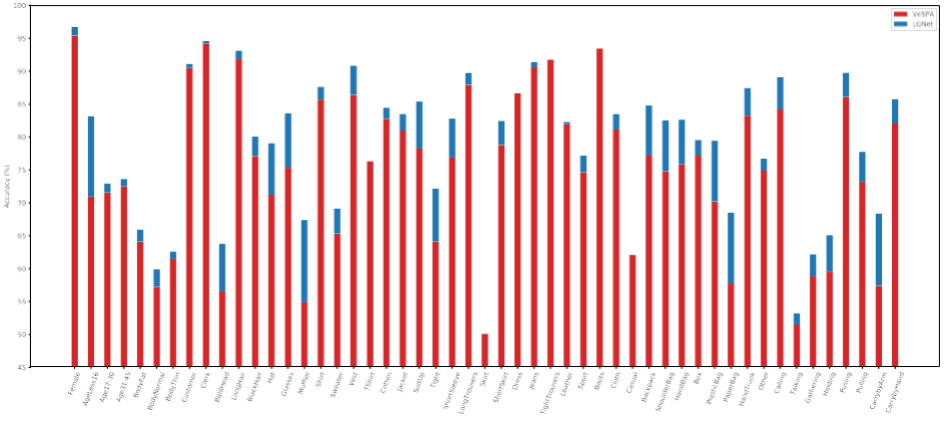

Figure 3: Mean accuracy comparison between our LG-Net and VeSPA on 51 attributes of RAP dataset. LG-Net significantly surpasses VeSPA in most attributes, especially attributes related to associated objects. The improvement in mean accuracy for object-type attributes proves the effectiveness of our designed localization guided architecture.

5 Ablation Study

To better demonstrate the effectiveness and advantage of our designed architecture, we apply component-wise ablation studies to explicitly address the contribution of each block.

Mapping Resolution. The original Inception-v2 produce feature maps in a mapping resolution of 7↛7. We stretch the feature maps into 14↛14 by changing the stride in *inception-4e*. The replacement leads to minor computational inefficiency, but the localization performance is apparently improved. Better localization result is helpful for our future studies on local feature extractions.

Table 2 shows the comparison between the result of the model with 14 ↛ 14 feature map and the 7 ↛ 7 baseline architecture. The result shows that network with 14 ↛ 14 feature map surpasses 7 ↛ 7 result by 0.92% in mA and 0.45% in Accuracy.

Dilated Convolution. Dilated convolution [33] is widely adopted in some recent-proposed networks for person re-identification [18] and some other tasks [10] to enlarge the receptive field. Dilated convolution is adopted in the *inception-4e* to maintain 14 ↛ 14 feature maps instead of directly replacing convolution stride by one. We compare the dilated convolution version with the ordinary convolution as shown in table 2. We compare dilated convolutions with dilation ratio 1, 2, and change dilation ratio to 2 on one branch only. We adopt the version with dilation ratio 1 and 2 for ultimate performance.

Affinity Strategy. For the fusion network, one key factor that affects the result is the proposal weighting strategy. We tried two strategies to calculate the affinity between bounding box extracted from Class Activation Map and Region of Interests. The adopted version is to calculate the Intersection over Union (IoU) as the weight of proposals. Apart from this version, we also tried to weight the proposal by directly counting the overlapping area of the corresponding bounding boxes. Table 2 demonstrates the effectiveness of using IoU as the weight. IoU brings 0.84% and 0.42% boost in mA and Accuracy respectively.

Localization Guidance. Localization Guidance Module is the core of our network. To

Dataset for Ablation Study	RAP				
Method	mA	Accu	Prec	Recall	F1
7 ↛ 7 Feature Map	75.43	65.94	**79.78**	77.05	78.39
14 ↛ 14 Feature Map	**76.35**	**66.39**	78.53	**78.98**	**78.75**
Dilation ratio = 1	76.35	66.39	78.53	78.98	78.75
Dilation ratio = 2	**77.05**	66.49	78.39	**79.33**	78.86
Dilation ratio = 1 and 2	76.79	**67.32**	**79.69**	79.28	**79.49**
Overlapping Area as Proposal Weight	77.84	67.58	80.22	79.32	79.77
IoU as Proposal Weight	**78.68**	**68.00**	**80.36**	**79.82**	**80.09**
LG-Net without localization	77.88	63.59	75.21	78.63	76.88
LG-Net	**78.68**	**68.00**	**80.36**	**79.82**	**80.09**

Table 2: Ablation Study: Effects of Components.

better illustrate the effectiveness of weight proposals by the affinity between proposals and localization boxes, we modify our network by removing the localization generator as well as the localization guidance module. All proposals are weighted evenly in the modified network. Table 2 shows the comparison between the modified LG-Net without localization guidance and the original LG-Net. Result suggests that localization module significantly improves the accuracy by 4.4%, which means that accurate localization of attributes is vital to the recognition of attributes for local feature extraction.

6 Conclusion

In this paper, we present a novel Localization Guided Network for the local feature extraction and utilization with attribute-specific localization guidance. Our LG-Net indicate that spatial information of attributes is helpful to the utilities of local features. Further ablation studies show the effectiveness of each component of our localization network and indicate that localization guidance is the key factor of success. Our network outputs the localization result parallel to the attribute predictions, and achieves the state-of-the-art results on the two largest pedestrian attribute datasets.

References

[1] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-taught object localization with deep networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[2] Archith John Bency, Heesung Kwon, Hyungtae Lee, S Karthikeyan, and BS Manjunath. Weakly supervised localization using deep feature maps. In *European Conference on Computer Vision*, pages 714–731. Springer, 2016.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[4] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. In *Proceedings BMVC 2014*, pages 1–12, 2014.

[5] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017.

[7] 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*, pages 789–792. ACM, 2014.

[8] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3): 275–293, 2012.

[9] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. Generative adversarial models for people attribute recognition in surveillance. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[10] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. *arXiv preprint arXiv:1712.04109*, 2017.

[11] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[14] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *Bmvc*, volume 2, page 8, 2012.

[15] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014.

[16] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 111–115. IEEE, 2015.

[17] 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.

[18] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017.

[19]Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. *arXiv preprint arXiv:1802.10171*, 2018.

[20]Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017.

[21]Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic, et al. Weakly supervised object recognition with convolutional neural networks. In *Proc. of NIPS*, 2014.

[22]Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[23]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.

[24]Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[25]M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017.

[26]Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[27]Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

[28]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 Workshops*, pages 87–95, 2015.

[29]Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision*, volume 2, 2017.

[30]Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2016.

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

[32] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[33]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.

[34]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*, pages 1637–1644, 2014.

[35]Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

[36]Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016.

[37]Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229, 2017.

[38]C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.