GSR-MAR: Global Super-Resolution for Person Multi-Attribute Recognition

Thomhert S. Siadari[1,2], Mikyong Han[2], and Hyunjin Yoon[1,2]
University of Science and Technology, Daejeon, Rep. of Korea[1]
Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea[2]
{thomhert,mkhan,hjyoon73}@etri.re.kr

Abstract

Person attribute recognition aims to predict attribute labels based on person's appearance usually captured from surveillance cameras. It is a challenging problem in computer vision due to poor imaging quality with complex background clutter and unconstrained viewing conditions from various angles and distances between person and surveillance cameras. In this paper, we address such a problem using an end-to-end network called Global Super-Resolution for Multi Attribute Recognition (GSR-MAR). GSR-MAR integrates a conversion process of low-resolution input images into high-resolution images and predicts person attributes from input images. Before performing the classification process, GSR-MAR not only converts low-resolution images to high-resolution images to recover details of image textures but also captures larger context information by using large separable convolutional layers. The experiment results on two popular benchmark datasets demonstrate the performance improvement and effectiveness of our GSR-MAR model over competing baselines.

1. Introduction

Person attribute recognition can be considered as a multi-label image classification based on full person body's appearance. It aims to predict multiple semantic attribute labels with binary outputs such as gender, age, carrying a backpack, carrying a bag, and wearing a hat. This task becomes highly demanding because of its benefits for various visual applications, such as video surveillance and image retrieval. However, in the real-world applications such as video surveillance, recognizing person attributes is challenging due to the poor quality of input images and unconstrained viewing conditions from varying angles and distances between a person and surveillance cameras. These visually degraded and obscured images can greatly affect the performance of person attribute recognition from real-world images.

Recently, deep learning-based methods have shown remarkable progress in person attribute recognition. Most of the methods used convolutional neural networks (CNNs) and its variants using the attention-based model [12], dividing the input image into several parts, and using human pose and body parts. Liu *et al*. [12] introduced a new attention-based deep neural network that captures multi-level features from different layers in the network in order to capture local and global features and then assemble the features to combine different semantic levels. Zhang *et al*. [19] explored a new model that first detects poselets of person and extracts each poselet using CNNs to get part-based deep representations. Li *et al*. [10] and Zhu *et al*. [20] divided an input image based on body parts and then extracted all parts using CNNs to get deep feature representations. However, these works only utilize different methods to extract features locally and globally by using different body parts or poselets. In contrast to existing methods, we aim to improve the recognition performance by feeding the classification network with high-resolution input images. However, acquiring a high-resolution image dataset is a time consuming and relatively expensive process. One intuitive solution is by converting low-resolution images into high-resolution images on-the-fly during training and inference processes [1].

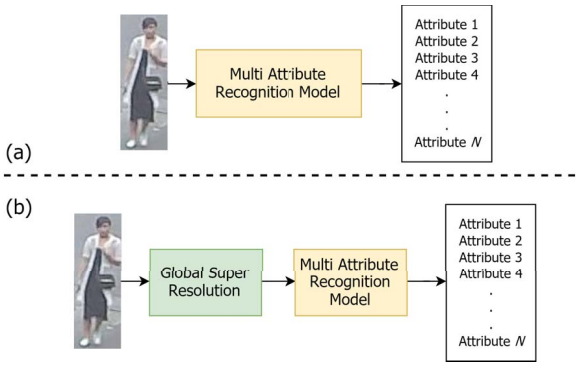


Figure 1: (a) Baseline, a deep learning-based method for person multi-attribute recognition. (b) Our proposed method that integrates global super-resolution before conducting multi-label classification in an end-to-end manner.

In this paper, we address such a problem using an end-to-end network called Global Super-Resolution for Multi Attribute Recognition (GSR-MAR). GSR-MAR integrates a conversion process of low-resolution input images into high-resolution images and predicts person attributes from the converted high-resolution features. GSR-MAR not only recovers details of textures from low-resolution images into high-resolution images but also captures larger context information before performing the classification process due to the use of large separable convolutional layers in its network. Figure 1 describes the difference between our proposed method and a baseline method. The baseline is a deep learning-based method for person multi-attribute recognition that does not consider resolution conversion process. We demonstrate the performance improvement and effectiveness of our GSR-MAR model over the baselines on two benchmark datasets RAP [11] and PA-100K [12]. The rest of the paper is organized as follows. Section II presents the related works in the field of person attribute recognition and super resolution. Section III describes the details of our proposed method. Section IV presents experiments and discusses our results. Section V concludes the paper.

2. Related Works

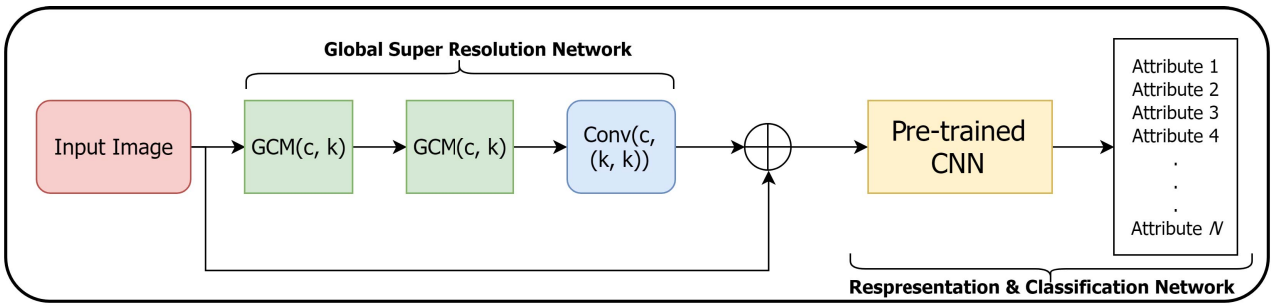
Person Attribute Recognition. Person attribute recognition is defined as a multi-label image classification task that has been widely applied to many applications such as person re-identification for video surveillance and image retrieval. Recently deep learning based methods significantly improved the performance of person multi attribute recognition [16, 12, 10, 18, 19, 20]. Sudowe *et al*. [16] introduced a CNN model that is trained by considering relationships and dependencies among all attributes in an end-to-end manner with independent loss for each attribute. Liu *et al*. [12] introduced a new attention-based deep neural network that captures multi-level features from different layers in order to capture local and global features and then assemble the features to form different semantic features. Li *et al*. [10] addressed the person multi-attribute recognition using two methods: deep learning based single attribute recognition (deepSAR) and deep learning multiple attribute recognition which considers imbalance samples from datasets. Wang *et al*. [18] formulated a model using recurrent neural network in order to exploit pedestrian attribute correlation from input image in sequential order. Zhang *et al*. [19] explored a new model that first detects poselets of person and extracts each poselet using CNNs to get part-based deep representation. Li *et al*. [10] and Zhu *et al*. [20] divided an input image based on body parts and then extracted all parts using CNNs to form deep feature representation. In contrast to our model, previous methods focus on learning different attribute-related features locally and globally by using different body parts or poselets but ignoring the fact that classification can be improved by using higher resolution image that can be converted along the way with classification process.

Super-Resolution. Earlier super-resolution methods that convert low-resolution into high-resolution images can be classified into four categories: image statistical methods, example-based methods, prediction model methods, and edge-based methods. Even though these methods are computationally fast, but they cannot build realistic image textures. In recent years, with the development of deep learning methods, several works have used CNN-based methods for super-resolution task which sets the state-of-the-art performance. It was started with Super-Resolution Convolutional Neural Network (SRCNN) by Dong *et al*. [4] that proposed CNNs to learn deep feature mapping for conversion of low-resolution images to high-resolution images. A major drawback of SRCNN is that it produces a higher computational complexity and visible artifacts. To address the the drawback of SRCNN and to accelerate the training and testing process, Dong *et al*. [5] proposed a new method called Fast Super-Resolution CNN (FSRCNN) that uses a compact CNN structure shaping like an hourglass. Shi *et al*. [15] proposed Efficient Sub-pixel Convolutional Networks (ESPCN) that only conduct up-scaling operation at the very last stage of its network. With the trend of deep learning methods that get deeper, Kim *et al*. [9] proposed a residual architecture to learn feature mapping for super-resolution purpose. Kim *et al*. [9] proposed a network with deep recursive layers in order to reduce network parameters. In this paper, we proposed to incorporate super-resolution sub-network to learn high resolution feature maps for multi-attribute recognition.

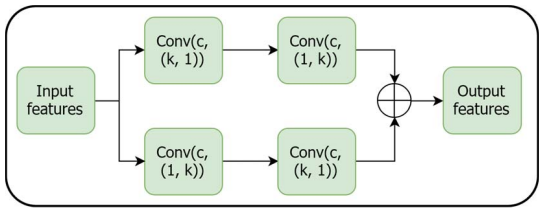
3. Proposed Method

3.1. Overview

The GSR-MAR integrates a conversion process of low-resolution input images into high-resolution images and predicts person attributes from input images. It consists of two sub-networks: Global Super Resolution (GSR) network and Representation & Classification (RC) network as shown in Figure 2(a). The GSR network converts the low-resolution images into high-resolution images, while RC network learns converted high-resolution feature maps and performs a multi-attribute classification. The multi-attribute recognition task is conducted as follows: given an input image I with corresponding N attribute labels, GSR-MAR receives a low-resolution input image I that will be converted into high-resolution image using GSR network. RC network then learns the output of GSR network and predicts the attribute scores in an end-to-end manner.



(a) Network architecture of GSR-MAR



(b) Global Convolutional Module (GCM)

Figure 2: (a) An overview of our GSR-MAR network. It consists of two sub networks: Global Super-Resolution Network (GSR) and Representation & Classification Network (RC). (b) The architecture of Global Convolutional Module (GCM) which is an essential module in our Global Super-Resolution Network.

3.2. Global Super Resolution Network

The main purpose of global super-resolution (GSR) network is to convert low-resolution images into high-resolution images in order to be fed into the following RC network. The GSR receives c and k parameters where c is number of output filters and k is the kernel size, respectively. Instead of using stacks of general convolutional layers, our GSR network consists of two Global Convolutional Modules (GCMs) and a convolutional layer as described in Figure 2(a). GCM employs large separable convolutional layers to capture larger context information from input feature maps as shown in Figure 2(b). The GCM is inspired from [13] that uses combination of large separable convolutional layers for semantic segmentation. The first large separable convolutional layers (upper line) consists of convolutional layers with kernel size of k×1 and 1×k, while the second separable convolutional layer (bottom line) consists of convolutional layers with 1×k and k×1. We then perform an element-wise addition to outputs of two separable convolutional layers. It should be noticed that all convolutional layers in GCM are employed with ReLU (Rectified Linear Unit) but the last convolutional layer in GSR is trained without activation function. After getting the output of GSR, we conduct a residual operation that will remove similarities between upsampled low-resolution images and converted high-resolution images.

3.3. Representation & Classification Network

The representation & classification (RC) network consists of several convolutional layers and fully-connected layers. The architecture of RC network is selected from various existing pre-trained models trained on ImageNet [14]. Specifically, we choose one model among five state-of-the-art models: ResNet50 [6], Xception [2], InceptionV3 [17], MobileNet [7], and DenseNet121 [8]. The pre-trained model without the last fully-connected layer is then followed by Global Average Pooling layer (GAP), a dropout layer, and a classifier. By using pre-trained model, we learn the transfer ability of network architectures trained on ImageNet [14].

3.4. Optimization

The GSR-MAR is trained in an end-to-end fashion using binary cross-entropy loss as in Eq. (1). Using this loss, we jointly learn the relationship of all attribute labels as formulated as follow:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} y_{ij} log(\hat{p}_{ij}) + (1-y_{ij}) log(1-\hat{p}_{ij}), \quad (1)$$

$$\hat{p}_{ij} = \frac{1}{1+exp(-x_{ij})}, \quad (2)$$

where $\hat{p}_{ij}$ is probability score for the j_{th} attribute of a input image I_i. y_{ij} is the true label that represents the j_{th} at-

Table 1: Performance of baseline method without fine-tuning of pre-trained model on RAP [11] and PA-100K [12] datasets.

Dataset	Backbone	mA	Accuracy	Precision	Recall	F1
RAP [11]	Xception [2]	49.98	36.16	79.83	38.81	52.29
	ResNet50 [6]	**50.03**	40.75	**75.49**	45.41	56.71
	InceptionV3 [17]	49.95	37.19	75.20	41.67	53.63
	MobileNet [7]	50.02	36.46	73.89	40.86	52.62
	DenseNet121 [8]	**50.03**	**40.96**	74.61	**46.07**	**56.96**
PA-100K [12]	Xception [2]	50.42	38.24	47.73	65.54	55.24
	ResNet50 [6]	**53.33**	**43.45**	**72.93**	**50.09**	**59.39**
	InceptionV3 [17]	50.14	40.83	65.37	49.39	56.26
	MobileNet [7]	50.23	37.53	67.76	44.24	53.53
	DenseNet121 [8]	51.07	41.91	67.84	50.06	57.61

Table 2: Performance of baseline method with fine-tuning of pre-trained model on RAP [11] and PA-100K [12] datasets.

Dataset	Backbone	mA	Accuracy	Precision	Recall	F1
RAP [11]	Xception [2]	50.06	39.88	73.05	45.69	56.22
	ResNet50 [6]	**54.72**	**47.78**	**78.88**	**53.33**	**63.61**
	InceptionV3 [17]	49.78	38.90	70.46	45.40	55.22
	MobileNet [7]	50.03	38.40	75.81	42.85	54.75
	DenseNet121 [8]	49.97	36.86	77.85	40.31	53.12
PA-100K [12]	Xception [2]	49.97	36.62	60.33	46.42	52.47
	ResNet50 [6]	**61.56**	**59.68**	**81.24**	**66.80**	**73.32**
	InceptionV3 [17]	50.40	39.19	62.87	47.45	54.08
	MobileNet [7]	51.84	37.64	58.58	48.72	53.19
	DenseNet121 [8]	49.98	29.38	52.31	37.83	43.91

tributes of a input image I_i, and x_i is the prediction results for an input image I_i.

4. Experiments

4.1. Dataset and Metrics

We perform our experiments on public benchmark person attribute datasets: (1) RAP [11] and PA-100K [12]. The Richly Pedestrian Dataset (RAP) [11] has a total of 41,585 images collected from 26 indoor surveillance cameras, in which every image in dataset has 72 attribute annotations. However, only 51 binary attributes with positive ratio more than 1% are selected for evaluations. It is then split into 33,268 images and 8,317 images for training and testing, respectively. During training, we select 20% of training images for validation data. PA-100K [12] has 100,000 images in total, split into 80,000 images for training data, 10,000 images for validation data, and 10,000 images for testing data. PA-100K has only 26 binary attribute annotations which is less than RAP dataset [11]. To evaluate the performance on these datasets, we follow the the settings of previous works that use label-based evaluation and sample-based evaluation method. Label-based or class-centric evaluation calculates the mean Accuracy (mA) by averaging the accuracy on the positive and negative samples. Sample-based or instance-centric evaluation computes accuracy, precision, recall, and F1 scores.

4.2. Implementation details

We formulate a baseline method which is basically the same as RC network. The baseline model is a pre-trained model (backbone) of ResNet50 [6], Xception [2], InceptionV3 [17], MobileNet [7], and DenseNet121 [8] based on ImageNet dataset [14] as its backbone. The input image for baseline and our proposed methods are low-resolution images, with data augmentation such as horizontal flip, rotation, width and height shift, and shear. Batch size of baseline model is 128, while batch size of our proposed method is 32. Furthermore, we use Stochastic Gradient Descent as the optimizer and apply polynomial decay to the learning rate with initial learning rate of 10^{-2}, the power of the polynomial of 1, momentum of 0.9, and maximum epoch of 50. The baseline and proposed methods are trained with the same settings to create a fair comparison. Both baseline and proposed methods are implemented using Keras library [3].

4.3. Results

In order to select the backbone model out of many pre-trained models to be used as RC network in our proposed method, we first conducted experiments using the baseline

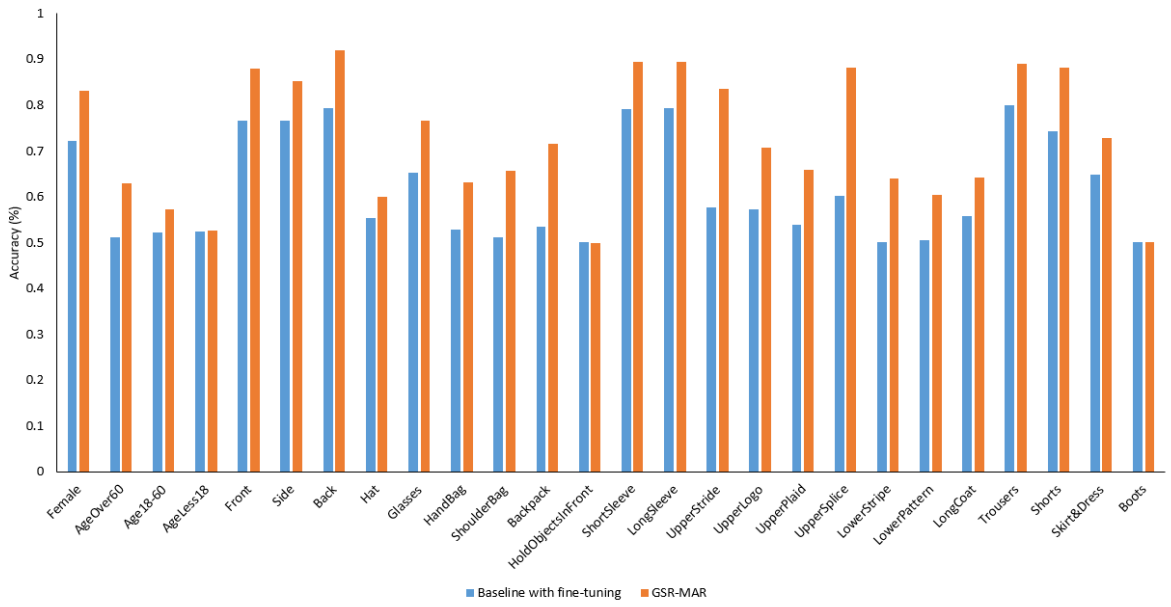


Figure 3: The accuracy comparison between baseline with fine-tuning and our proposed method on 26 attributes of PA-100K dataset.

Table 3: Quantitative comparison of GSR-MAR over the baselines on RAP [11] and PA-100K [12] datasets.

Dataset	Model	mA	Accuracy	Precision	Recall	F1
RAP [11]	Baseline - w/o finetuning	50.03	38.4	75.809	42.847	54.75
	Baseline - w/ finetuning	54.723	47.778	78.883	53.331	63.609
	GSR-MAR	**67.76**	**63.44**	**82.27**	**71.82**	**76.69**
PA-100K [12]	Baseline - w/o finetuning	50.01	40.746	75.488	45.409	56.707
	Baseline - w/ finetuning	61.56	59.68	81.24	66.8	73.32
	GSR-MAR	**72.43**	**73.46**	**87.68**	**79.94**	**83.63**

method on both RAP [11] and PA-100K. Furthermore, we also investigate the transfer ability of pre-trained models trained on ImageNet [14] for person attribute recognition task. Specifically, we conduct a study of transfer learning across five modern deep learning classification models as mentioned on Section 4.2. We measure transfer learning performance in two settings (1) fixed feature extractors without fine-tuning and only change the last classification layer, (2) fine-tuning using weights initialization from ImageNet [14] for all network layers. Experiment results of these two settings are reported in Table 1 and Table 2, respectively. We used a similar setting without an excessive hyper-parameter tuning for creating a fair comparison. We show that the best performing models on RAP dataset without fine-tuning are ResNet50 [6] and DenseNet121 [8] with 50.03% in terms of mA. However, the best performing model on RAP dataset [11] with fine-tuning is ResNet50 [6] with 54.72% in terms of mA. Fine-tuning process on RAP dataset [11] fails to bring improvement for most models. While on PA-100K dataset [12], fine-tuning process significantly improves the performance for ResNet50 [6], slightly improves the MobileNet [7] performance, and fails to bring any improvement for the other models. Based on these experiments, we decided to use ResNet50 [6] for our RC network.

We evaluate our proposed method over the baseline method with or without fine-tuning process. GSR-MAR outperforms the baseline with 67.76% and 72.43% in terms of mA for RAP [11] and PA-100K datasets [12], respectively. Using GSR-MAR, the performance is increased from 54.72% to 67.76% on RAP dataset [11] and from 61.56% to 72.43% on PA-100K [12] in terms of mA. In addition, our GSR-MAR also outperforms the baseline in terms of accuracy, precision, recall, and F1 score as reported in Table 3. Moreover, we also provide detailed comparison in term of accuracy score between baseline with fine-tuning and our

proposed method on 26 attributes of PA-100K dataset [12] as seen in Figure 3. GSR-MAR surpasses the baseline in most attributes. This result demonstrates that GSR-MAR can increase the overall performances.

5. Conclusion

In this paper, we proposed a new multi-attribute recognition method called GSR-MAR to recognize person multi-attributes. GSR-MAR combines a global super-resolution network and classification network. Hence, it not only recovers the detail texture of low-resolution input images into high-resolution images, but also captures larger context information from input images before performing the classification process which results in recognition performance improvement. The proposed method is evaluated over two public benchmark datasets. The experiment results demonstrate the effectiveness of our methods compare to the baseline methods in all performance metrics.

Acknowledgments

This work was supported by The Cross-Ministry Giga KOREA Project grant of the Korea government (MSIT), Rep. of Korea (No.GK19P0600, Development and Demonstration of Smart City Service over 5G Network).

References

[1] D. Cai, K. Chen, Y. Qian, and J. K. Kämäräinen. Convolutional low-resolution fine-grained classification. *Pattern Recognition Letters*, 119:166–171, 2019.

[2] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[3] F. Chollet et al. Keras. `https://keras.io`, 2015.

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, feb 2016.

[5] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9906 LNCS:391–407, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[9] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:1637–1645, 2016.

[10] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-Aware features over body and latent parts for person re-identification. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:7398–7407, 2017.

[11] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A Richly Annotated Dataset for Pedestrian Attribute Recognition. 2016.

[12] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October(c):350–359, 2017.

[13] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[15] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[16] P. Sudowe, H. Spitzer, and B. Leibe. Person Attribute Recognition with a Jointly-Trained Holistic CNN Model. *Proceedings of the IEEE International Conference on Computer Vision*, 2016-February:329–337, 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[18] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute Recognition by Joint Recurrent Learning of Context and Correlation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:531–540, 2017.

[19] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[20] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229, 2017.