Generative Adversarial Models for People Attribute Recognition in Surveillance

Matteo Fabbri Simone Calderara Rita Cucchiara

University of Modena and Reggio Emilia
via Vivarelli 10 Modena 41125 Italy
name.surname@unimore.it

Abstract

In this paper we propose a deep architecture for detecting people attributes (e.g. gender, race, clothing ...) in surveillance contexts. Our proposal explicitly deal with poor resolution and occlusion issues that often occur in surveillance footages by enhancing the images by means of Deep Convolutional Generative Adversarial Networks (DCGAN). Experiments show that by combining both our Generative Reconstruction and Deep Attribute Classification Network we can effectively extract attributes even when resolution is poor and in presence of strong occlusions up to 80% of the whole person figure.

1. Introduction

Surveillance cameras has spread rapidly in most of the cities all around the world. Among the surveillance tasks computer vision has focused on tracking and detection of targets where targets are described by their visual appearance. Nevertheless, recently the task of capturing as many people characteristics as possible has gained importance for better understanding a situation and its attendants. The task, referred in literature as *attribute recognition* [13], consists in detecting people attributes (such as age, sex, etc.) and items (backpacks, bags, etc.) of people through security cameras. While this task have been profitably attacked from a face recognition perspective capturing gender, age, and race,[8, 6], very few works focus on whole people body. Among these, most of them, [1, 12], consider people always unoccluded and at full resolution that is not the case when dealing with surveillance footages. In fact, surveillance cameras, that have typically a far field of view, are massively affected by resolution issues and people occlusion, Fig. 1.

In this work we propose an attribute recognition method that explicitly deals with resolution and occlusions by exploiting a generative deep network approach [15]. Our proposal consists of three deep networks. The first classifies people attributes given full body images. The others focus on enhancing the input image by raising its resolution and trying to reconstruct images from occlusion by means of a generative convolutional approach [10]. To our knowledge, this is the first work that considers this task in a surveillance context by explicitly dealing with those both issues.

Figure 1. Resolution and occlusions issues and reconstructed frames by our generative approach.

2. Related Work

Early works on attribute recognition usually treat attributes by independently training a different classifier for each attribute, [20, 2]. More recently Convolutional Neural Networks (CNN), enable researchers to mine the relationship between attributes and are preferred on large scale object recognition problems because of their advanced performances. There are large bodies of work on CNNs, like [8] which undertakes the task of occlusion and low-resolution robust facial gender classification, or [6, 19] that predict facial attributes from faces in the wild. Many other works like [11, 17] propose different methods to achieve attribute classification like gender, smile and age in an unconstrained environment. However, those technique involve only facial

978-1-5386-2939-0/17/$31.00 ©2017 IEEE

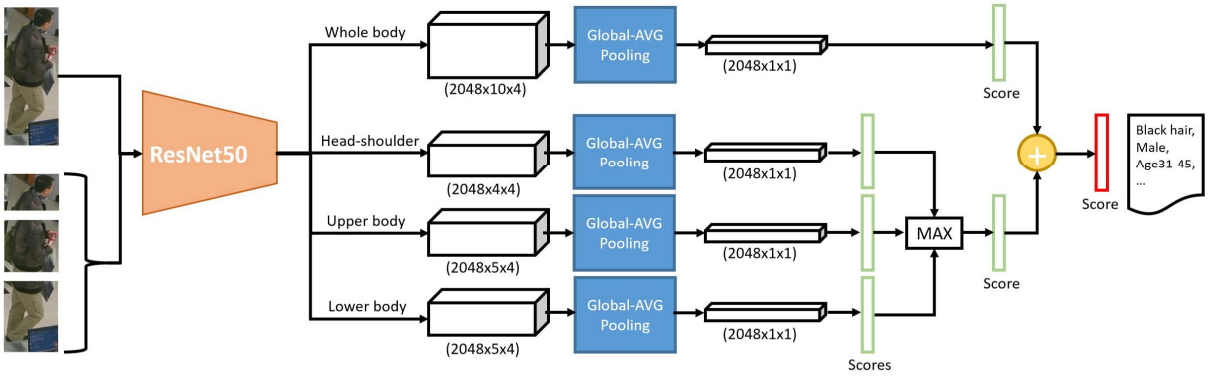

Figure 2. Architecture of our Attribute Classification Network.

images and are not suitable for surveillance tasks. Moreover [1] addresses the problem of describing people based on clothing attributes. Nevertheless, our work considers the person as a whole and does not focus only on clothing classification. More recent works that rely on full-body images to infer human attributes are the Attribute Convolutional Net (ACN) and Deep learning based Multi-Attribute joint Recognition model (DeepMAR), [16, 12]. ACN jointly learns different attributes through a jointly-trained holistic CNN model, while DeepMAR utilizes the prior knowledge in the object topology for attribute recognition. In [18, 4] attributes classification is accomplished by combining part-based models and deep learning by training pose-normalized CNNs. Additionally, MLCNN [21] splits the human body in 15 parts and train a CNN for each of them while DeepMAR* [13] divides the whole body in three parts which correspond to the headshoulder part, upper body and lower body of a pedestrian respectively. Furthermore, [14] tackles the problem of attribute recognition by improving a part-based method within a deep hierarchical context. Nevertheless, the majority of those methods relies on high resolution images and does not encompass the problem of occlusion. Recent works on image super-resolution exploit Generative Adversarial Networks (GAN) [5], and more precisely Deep Convolutional Adversarial Networks (DCGAN) [15], in order to generate high resolution images starting from low resolution ones [3, 10].

3. Method

The contribution of this work consists on three networks: a baseline state-of-the-art part-based architecture for human attribute classification based on ResNet [7], a generative model that aims to reconstruct the missing body parts of people in occluded images and a second generative model that is capable of enhance the resolution of images at low resolution.

3.1. Attribute Classification Network

The proposed approach for human attribute classification is inspired by the previous part-based works thus capable of learning pose-normalized deep feature representations from different parts. By blending together the capability of neural networks with a part-based approach grants robustness when facing unconstrained images dominated by the effects of pose and viewpoints. Inspired by [14], we propose to decompose the input image I into blocks which correspond to the whole body b and a set of parts $\{p \in P\}$, (inputs in Fig 2). We choose three parts: head-shoulder section, upper body and lower body of the pedestrian. Those four blocks are then passed through the ResNet50 network [7] pretrained on ImageNet classification [9] to obtain four part-based convolutional feature maps. Note that, in order to achieve this, we replaced the last average pooling 7×7 in ResNet50 with a global average pooling. This allowed us to fed the network with images that have one of their dimension smaller than 227. After computing the feature maps, we branch out two attribute score paths. On the first path, we use the full body feature maps in order to obtain a prediction score based on the whole person. We incorporate the part-based technique in the second path where we compute a prediction score for each image partition (scores in Fig 2), followed by a $\max$ score operation that aims to select the most descriptive part for each attribute. This operation is needed because human attributes often reside in a specific body area (e.g. "hat" is in the head-shoulder part). The final attribute prediction is performed by adding the whole body score to the part score:

$$\begin{aligned} Score_i(I) &= Score_i(b) + Score_i(P) \\ &= w_{i,b}^T \cdot \phi(b) + \max_{p \in P} w_{i,p}^T \cdot \phi(p) \end{aligned} \quad (1)$$

where $w_{i,\cdot}$ are the scoring weights of the ith attribute for different regions, while $\phi(\cdot)$ are the feature maps from different regions.

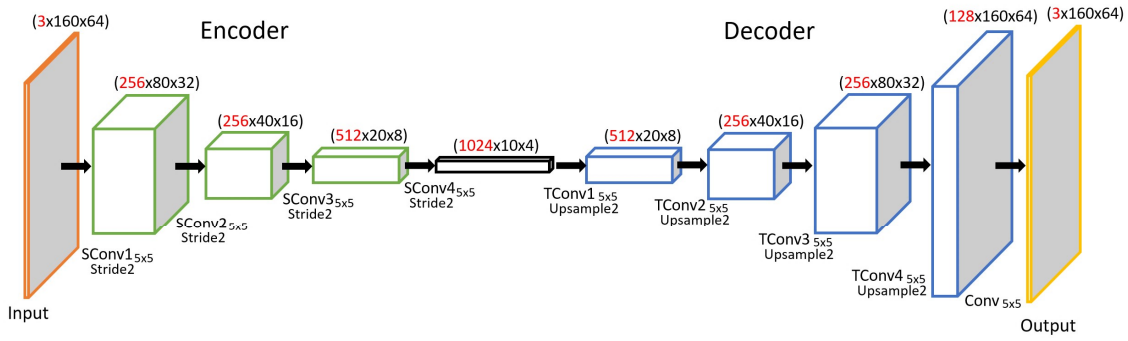

Figure 3. Architecture of our Reconstruction Network.

The whole network, depicted in Fig. 2, is trained using a *weighted binary cross entropy* loss with a different weight for each attribute. The need of using a weighted loss arise from the fact that the distributions of attributes classes in the dataset may be imbalanced:

$$Loss_C = -\sum_{i=1}^{A} \frac{1}{2r_i} \cdot y_i \cdot \log(\hat{y_i}) + \frac{1}{2(1-r_i)} \cdot (1-y_i) \cdot \log(1-\hat{y_i}) \quad (2)$$

where A is the total number of attributes, y is the ground truth label vector and $\hat{y}$ is the predicted label vector. For each attribute, r is the ratio between the number of samples that hold that attribute and the total number of samples.

3.2. Reconstruction Network

Our second goal consist in making the previous architecture robust to occlusion integrating a system capable of removing the obstructions and replacing them with body parts that could likely belong to the occluded person. In the worst case scenario, even though the replaced body parts does not reflect the real attributes of the person, the reconstruction still helps the classifier: by removing the occlusion we produce an image that contains only the subject without noise that could lead to misclassifications. For example, an image containing a person occluded by another person could induce the network to classify the attributes of the person in the foreground which is not the subject of the image. In order to accomplish this, we train a generative function G^R capable of estimating a reconstructed image I^R from an occluded input image I^O. During training I^O is obtained by artificially partially overlapping an image I with another image. To achieve our goal we train a generator network as a feed-forward CNN $G^R_{\theta_g}$ with parameters θ_g. For N training images we solve:

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{n=1}^{N} Loss_R(G^R_{\theta_g}(I^O_n), I_n) \quad (3)$$

Here $\hat{\theta}_g$ is obtained by minimizing the loss function $Loss_R$ described at the end of this section.

Following [5], we further define a discriminator network $D^R_{\theta_d}$ with parameters θ_d that we train alongside with $G^R_{\theta_g}$ in order to solve the adversarial min-max problem:

$$\min_{G^R} \max_{D^R} \mathbb{E}_{I\sim p_{data}(I)}[\log D^R(I)] + \mathbb{E}_{I^O\sim p_{gen}(I^O)}[\log 1 - D^R(G^R(I^O))] \quad (4)$$

The purpose of the discriminator D^R is to distinguish generated images from real images, meanwhile the generator G^R is trained with the aim of fooling the discriminator D^R. With this approach we obtain a generator model capable of learning solution that are similar to not occluded images thus indistinguishable by the discriminator. Inspired by [15] we propose the generator's architecture illustrated in Figure 3. Specifically, in the encoder we use four strided convolutional layers (with stride 2) to reduce the image resolution each time the number of feature is doubled, SConvs in Fig. 3. The decoding uses four transposed convolutional layers (also known as fractionally strided convolutional layers) to increase the resolution each time the number of feature is halved, and a final convolution, TConvs in Fig. 3. We use Leaky ReLU as activation function in the encoding phase and ReLU in the decoding phase. We adopt batch-normalization layers before activations (except for the last Conv) and a kernel size 5×5 at each step. The discriminator architecture is similar to the generator's encoder except for the number of filter, which increase by a factor of 2 from 128 to 1024. The resulting 1024 feature maps are followed by one sigmoid activation function in order to obtain a probability useful for the classification problem. We use batch-normalization before every Leaky ReLU activation, except for the first layer.

The definition of the loss function $Loss_R$ is fundamental for the effectiveness of our generator network. Borrowing the idea from [10], we propose a loss composed by a

weighted combination of two components:

$$Loss_R = Loss_{SSE} + \lambda Loss_{gen} \quad (5)$$

Here $Loss_{SSE}$ is the reconstruction loss based on sum of squared errors of prediction (SSE) which let the generator predict images that are pixel-wise similar to the target image. The pixel-wise SSE is calculated between downsized versions of the generated and target images, first applying an averaged pooling layer. This is because we want to avoid the standard "blurred" effect that MSE and SSE trained autoencoders suffer from. In our experiments we used a λ equals to 10^{-1}.

The second component $Loss_{gen}$ is the actual adversarial loss of the generator G^R which encourages the network to generate perceptually good solutions that are in the subspace of person-like images. The loss is defined as follows:

$$Loss_{gen} = \sum_{i=1}^{N} \log(1 - D^R(G^R(I^O))) \quad (6)$$

Where $D^R(G^R(I^O))$ is the probability of the discriminator labeling the generated image $G^R(I^O)$ as being a real image.

3.3. Super Resolution Network

Our last goal is to integrate our system with a network capable of enhancing the quality of images that have poor resolution. This task is accomplished by training another generative function G^S capable of estimating an high resolution image I^H from a low resolution input image I^L. During training I^L is obtained from the original image I by performing a simple downsample operation with factor $r = 4$. To achieve our goal we train the generator network as a feed-forward CNN likewise we did for G^R. As for the Reconstruction Network, we define the discriminator D^S in the same way we defined D^R. For D^S we used the same architecture used in D^R. The architectural differences between the two models reside in the number of layers: in the Super Resolution Network we used three strided convolution (in the Encoder), with 256, 512 and 1024 features respectively and five transposed convolutions (in the Decoder) that follow the pattern 512, 256, 256, 128, 128. The motivation is that the input image in G^S is two time smaller with respect to the input image in G^R. Moreover in D^S we used five strided convolution with the number of filter that increase by a factor of 2 from 128 to 2048. Eventually we set the λ value used to weight the loss components to 1.

4. Experiments

We conduct our experiments using the new RAP [13] dataset, a very richly annotated dataset with 41,585 pedestrian samples, each of which is annotated with 72 attributes as well as viewpoints, occlusions and body parts information. As recommended by [13], we measure performances with five metrics: mean accuracy (mA), accuracy, precision, recall and F1 score. We perform three type of experiments. First, we comparative evaluated the performances of our Attribute Classification Network comparing the results with other deep SoA approaches. Secondly, we corrupted the dataset with occlusions (occRAP 2^{nd} row in Fig. 4)and tested the benefits achieved by the combination of our Reconstruction Network, Sec. 3.2 with the Attribute Classification one. Thirdly, we corrupt the dataset by lowering the resolution (lowRAP 3^{rd} row in Fig 4) and evaluate the contribution of our Superesolution Network, Sec. 3.3, in conjuction with Attribute Classification. Eventually, we propose a complete classification pipeline where all the three network are combined together on both low-res and occluded images. The occRAP dataset is produced by randomly overlapping RAP images in order to artificially reproduce the occlusions. Note that in our experiment we focused the attention only on one type of occlusion: the occlusion that cover the bottom part of an image where the occluded portion have been randomly sampled from 50% to 80% occlusion rate. lowRap, instead, is obtained by performing a simple downsample with factor 4 from the original RAP images.

Figure 4. First row original RAP dataset. Second row occlusion RAP dataset where the images have been occluded from 50 to 80% by pasting the upper body of another subject. Third row low resolution RAP where the image are downsamples of a factor 4 on width and height respectively.

Attribute Classification Following [13] we conducted the experiments on the RAP dataset with 5 random splits. For each split, totally 33,268 images are used for training and the rest 8,317 images are used for testing. Due to the unbalanced distribution of attributes in RAP we selected the 50 attributes that have the positive example ratio in the dataset higher than 0.01. For each image we also add one attribute corresponding to the occlusion of our interest (*occlusion down* attribute). For each mini-batch, we resized the images to a fixed dimension of 320×128. In order to

Method	mA	Accuracy	Precision	Recall	F1
ACN [16]	69.66	62.61	**80.12**	72.26	75.98
DeepMAR [12]	73.79	62.02	74.92	76.21	75.56
DeepMAR* [13]	74.44	63.67	76.53	77.47	77.00
Our	**79.73**	**83.97**	76.96	**78.72**	**77.83**

Table 1. Comparison with SoA on the RAP dataset.

split the figure in the three parts we divide the height in 10 blocks and pick the top 4 for the head-shoulder part, the third to the seventh for the upper body part and the sixth to the tenth for the lower body part. The network is trained using stochastic gradient descent with a learning rate 10^{-5}, learning rate decay 10^{-6} and momentum 0.9. We used 8 images per mini-batch. Tab. 1 shows the results on RAP dataset where our baseline is compared against state-of-the-art methods on the same 51 attributes [1]. It can be shown that our network perform favourably in terms of Accuracy being competitive in both Precision and Recall related metrics. This is mainly due to the adoption of the fixed body part partitions of the image that, in case of people images from surveillance cameras, represent a reliable partition of the body in its parts. Additionally, parts scoring maximization allow for selecting the most reliable score for every individual attribute thus increasing the classification accuracy.

Reconstruction We trained our Reconstruction Network with the occRAP training set and simultaneously providing the network with the original not occluded images associated to the inputs in order to compute the $Loss_{SSE}$. For optimization we used Adam with $\beta_1 = 0.5$ and learning rate of 0.002. We alternate updates to the discriminator and generator network with $K = 1$ as recommended in [5].
Furthermore, the aim is to quantify the impact that occluded images have in the classification task. We firstly fed our classification network with images picked from occRAP testing set obtaining the results reported in Tab.2 while visual examples are depicted in Fig. 5. Secondly we repeated the experiment manipulating the input images using our Reconstruction Network with the aim of removing the occlusion. In the same table are reported the results that shows a significant improvement. From the results, it emerges that our Reconstruction Network provide a reasonable guess of the occluded person appearance being able of learning from the visible part a potentially useful image completion.

Resolution The dataset used for training the Super Resolution Network is the lowRAP. We trained the network with the original size images associated to the inputs in order to compute the $Loss_{SSE}$. We adopted the same optimizer and the same K value used for the Reconstruiction Network. In order to evaluate how the low resolution affects the performance on attributes classification, we inputed the lowRAP test set to our Attribute Classification Network, after a 4x bilinear upsampling. Subsequently, we passed the images through our Super Resolution Network before attribute classification. As can be seen in Tab. 2 the adoption of our Super Resolution network leads to an important improvement being able to keep more information w.r.t. the upsampling.

[1] Complete per-attribute results are in the supplementary material.

Input	mA	Accuracy	Precision	Recall	F1
occRap experiment					
occRAP	57.70	61.00	33.26	41.63	33.25
occRAP + NET	**68.81**	**74.54**	**57.29**	**58.91**	**58.09**
lowRap experiment					
lowRAP	63.80	74.51	44.47	49.56	40.37
lowRAP + NET	**76.02**	**80.12**	**69.56**	**73.12**	**71.30**
Complete Experiment					
Corrupted	60.68	72.75	38.67	45.47	41.80
Restored	**65.82**	**76.01**	**48.98**	**55.50**	**52.04**

Table 2. Experiments with occlusions (occRAP experiment), low resolution (lowRAP experiment) and Complete Model. The complete experiment uses the merge of test sets occRAP and lowRAP. *Corrupted* are the score of the Attribute Classification network on plain input data. *Restored* are the results when using our complete pipeline.

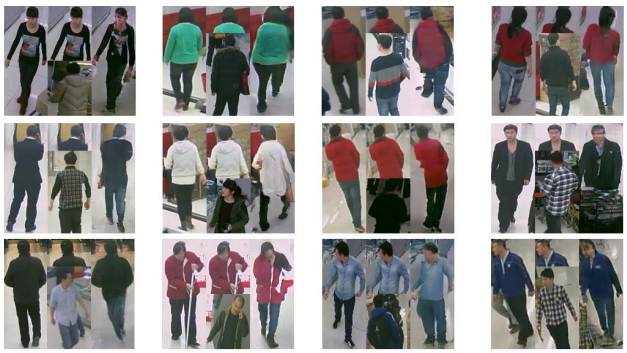

Figure 5. Visual examples of Generative Reconstruction. For every triplet of images: (Leftmost)the original image; (Middle)Occluded input; (Rightmost)Reconstruction/Guessed unoccluded image.

Complete Model Our final experiment consists in testing all our networks in order to build a system that is able to detect corrupted images and consequently react performing a restoring operation when possible. To achieve this we propose a simple algorithm where the input is passed through the Super Resolution Network only if the input image is smaller than the network input. The image is then passed through the Classification Network and, if the *occlusion down* attribute is positively triggered (the test F1 score of the *occlusion down* attribute is $> 85\%$), the image is passed through the Reconstruction Network. The reconstructed image is finally fed again to the Classification Net-

work to output the final scores. The test were performed on the merge of occRAP and lowRAP test sets. Tab. 2 highlights the improved results obtained by this pipeline w.r.t. our Deep Attribute Classification Network alone.

5. Conclusions

In this work we presented the use of Deep Generative Network for image enhancing in people attributes classification. Our Generative Network have been designed to overcome two common problems in surveillance scenarios, namely people resolution and occlusions. Experiments have shown that jointly enhancing images before feeding them to an attribute classification network can improve the results even when input images are affected by those issues. In further works we will explore the fusion of the networks in a single end-to-end model that can automatically choose which enhancement network activates by looking at images at test time. We find this line of work can foster research about the problem of attribute classification in surveillance contexts where camera resolution and positioning cannot be neglected.

6. Acknowledgements

The work is supported by the Italian MIUR, Ministry of Education, Universities and Research, under the project COSMOS Prin 2015 programme.

References

[1] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015.

[2] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*, 2015.

[3] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015.

[4] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *IEEE International Conference on Computer Vision*, pages 1080–1088, 2015.

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

[6] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, and M. Savvides. Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–77, 2016.

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

[10] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[11] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

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

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

[14] Y. Li, C. Huang, C. C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *IEEE International Conference on Computer Vision*, pages 684–700. Springer, 2016.

[15] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

[17] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–38, 2016.

[18] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[19] Y. Zhong, J. Sullivan, and H. Li. Leveraging mid-level deep representations for predicting face attributes in the wild. In *IEEE International Conference on Image Processing*, pages 3239–3243, 2016.

[20] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *IEEE International Conference on Computer Vision Workshops*, pages 331–338, 2013.

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