

Deep Learning Face Attributes in the Wild

Ziwei Liu^{1,3} Ping Luo^{3,1} Xiaogang Wang^{2,3} Xiaoou Tang^{1,3}

¹Department of Information Engineering, The Chinese University of Hong Kong

²Department of Electronic Engineering, The Chinese University of Hong Kong

³Shenzhen Key Lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China

{lzw13, pluo, xtang}@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

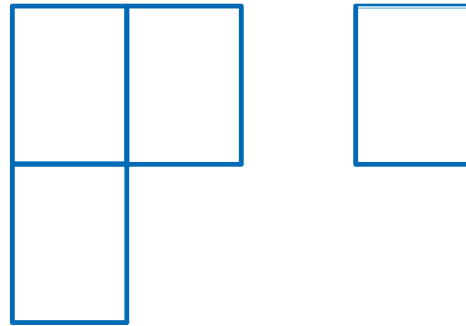
Predicting face attributes in the wild is challenging due to complex face variations. We propose a novel deep learning framework for attribute prediction in the wild. It cascades two CNNs, LNet and ANet, which are fine-tuned jointly with attribute tags, but pre-trained differently. LNet is pre-trained by massive general object categories for face localization, while ANet is pre-trained by massive face identities for attribute prediction. This framework not only outperforms the state-of-the-art with a large margin, but also reveals valuable facts on learning face representation.

(1) It shows how the performances of face localization (LNet) and attribute prediction (ANet) can be improved by different pre-training strategies. (2) It reveals that although the filters of LNet are fine-tuned only with image-level attribute tags, their response maps over entire images have strong indication of face locations. This fact enables training LNet for face localization with only image-level annotations, but without face bounding boxes or landmarks, which are required by all attribute recognition works. (3) It also demonstrates that the high-level hidden neurons of ANet automatically discover semantic concepts after pre-training with massive face identities, and such concepts are significantly enriched after fine-tuning with attribute tags. Each attribute can be well explained with a sparse linear combination of these concepts.

1. Introduction

Face attributes are beneficial for multiple applications such as face verification [15, 2, 24], identification [20], and retrieval. Predicting face attributes from images in the wild is challenging, because of complex face variations such as poses, lightings, and occlusions as shown in Fig.1.

Attribute recognition methods are generally categorized into two groups: global and local methods. Global methods extract features from the entire object, where accurate



This work revisits global methods by proposing a novel deep learning framework, which integrates two CNNs, LNet and ANet, where LNet locates the entire face region and ANet extracts high-level face representation from the located region. The novelties are in three aspects. Firstly, LNet is trained in a weakly supervised manner, i.e. only image-level attribute tags of training images are provided, making data preparation much easier. This is different from training face and landmark detectors, where face bounding boxes and landmark positions are required. LNet is pre-trained by classifying massive general object categories, such that its pre-trained features have good generalization capability on handling large background clutters. LNet is then fine-tuned by attributes tags. We demonstrate that features learned in this way are effective for face localization and also can distinguish subtle differences between human faces and analogous patterns, such as a cat face.

Secondly, ANet extracts discriminative face representation, making attribute recognition from the entire face region possible. ANet is pre-trained by classifying massive face identities and is fine-tuned by attributes. We show that the pre-training step enables ANet to account for complex variations in the unconstrained face images.

Thirdly, within the rough locations of face regions provided by LNet, averaging the predictions of multiple patches can improve the performance. A simple way is to evaluate the feed-forward pass for each single patch. However, it is slow and has a lot of redundant computation. A novel fast feed-forward scheme is proposed to replace patch-by-patch evaluation. It evaluates images with arbitrary sizes with only one-pass feed-forward operation. It becomes non-trivial if the filters are locally shared, while studies [27, 26] showed that locally shared filters perform better in face related tasks. This is solved by proposing an interleaved operation.

Besides proposing new methods, our framework also reveals valuable facts on learning face representation. They not only motivate this work but also benefit future research on face and deep learning. (1) It shows how pre-training with massive object categories and massive identities can improve feature learning for face localization and attribute recognition, respectively. (2) It demonstrates that although filters of LNet are fine-tuned by attribute tags, their response maps over the entire image have strong indication of face location. Good features for face localization should be able to capture rich face variations, and more supervised information on these variations improves the learning process. The examples in Fig. 1 (a) show that as the number of attributes decreases, the localization capability of learned neurons gets reduced dramatically. (3) ANet is pre-trained with massive face identities. It discloses that the pre-trained high-level hidden neurons of ANet implicitly learn and discover semantic concepts that are related to identity, such

as race, gender, and age. It indicates that when a deep model is pre-trained for face recognition, it implicitly learns attributes. The performance of attribute prediction drops without this pre-training stage.

The main contributions are summarized as follows. (1) We propose a novel deep learning framework, which combines massive objects and massive identities to pre-train two CNNs for face localization and attribute prediction, respectively. It achieves state-of-the-art attribute classification results on both the challenging CelebFaces [26] and LFW [12] datasets, improving existing methods by 8 and 13 percent, respectively. (2) A novel fast feed-forward algorithm for CNN with locally shared filters is devised. (3) Our study reveals multiple valuable facts on learning face representation by deep models. (4) We also contribute a large facial attribute database with more than eight million attribute labels and it is 20 times larger than the largest publicly available dataset.

1.1. Related Work

Extracting hand-crafted features at pre-defined landmarks has become a standard step in attribute recognition [9, 15, 4, 2]. Kumar et al. [15] extracted HOG-like features on various face regions to tackle attribute classification and face verification. To improve the discriminativeness of hand-crafted features given a specific task, Bourdev et al. [4] built a three-level SVM system to extract higher-level information. Deep learning [18, 34, 23, 7, 19, 32, 31, 13, 33, 22, 3, 28] recently achieved great success in attribute prediction, due to their ability to learn compact and discriminative features. Razavian et al. [23] and Donahue et al. [7] demonstrated that off-the-shelf features learned by CNN of ImageNet [13] can be effectively adapted to attribute classification. Zhang et al. [32] showed that better performance can be achieved by ensembling learned features of multiple pose-normalized CNNs. The main drawback of these methods is that they rely on accurate landmark detection and pose estimation in both training and testing steps. Even though a recent work [31] can perform automatic part localization during test, it still requires landmark annotations of the training data.

2. Our Approach

Framework Overview Fig.2 illustrates our pipeline where LNet locates the entire face region in a coarse-to-fine manner as shown in (a) and (b), while ANet extracts features for attribute recognition as shown in (c).

Different from existing works that rely on accurate face and landmark annotations, LNet is trained in a weakly supervised manner with only image-level annotations. Specifically, it is pre-trained with one thousand object categories of ImageNet [6] and fine-tuned by image-level attribute tags. The former step accounts for background



Nipgct UXO

"

Uo kipi



Figure 3. (a.1) Original image. (a.2)-(a.4) are averaged response maps in C5 of LNet₀ after pre-training (a.2), fine-tuning (a.3) and directly training from scratch with attribute tags but without pre-training (a.4). (b) Determine threshold.

ground. To determine the threshold, we select 2000 images, each of which contains a single face, and 2000 background images from SUN dataset [29]. For each image, EdgeBox [35] is adopted to propose 500 candidate windows, each of which is measured by a score that sums over its response values normalized by its window size. A larger score indicates the localized pattern is more likely to be a face. Each image is then represented by the maximum score over all its windows. In Fig 3 (b), the histogram of the maximum scores shows that these scores clearly separate face images from background images. The threshold is chosen as the decision boundary as shown in Fig 3 (b). More results are given in Fig 6 (a), showing that the above strategy can precisely localize face within a single test image. Since each training image only contains one single face, we localize a face region using the window with the largest score during training.

To understand why rich attribute information enables accurate face localization, one could consider the examples in Fig.4. If only a single detector [17, 21] is used to classify all the positive and negative samples in Fig.4 (a), it is difficult to handle complex face variations. Therefore, multi-view face detectors [30] were developed in Fig.4 (b), i.e. face images in different views are handled by different detectors. View labels were used in training detectors and the whole training set is divided into subsets according to views. If views are treated as one type of face attributes, learning face representation by predicting attributes with deep models actually extends this idea to extreme. As shown in Fig 4 (c), a filter (or a group of filters) functions as a detector of an attribute. When a subset of neurons are activated, they indicate the existence of face images with a particular attribute configuration. The neurons at different layers can form many activation patterns, implying that the whole set of face images can be divided into many subsets based on attribute configurations, and each activation pattern corresponds to one subset (e.g. ‘pointy nose’, ‘rosy cheek’, and ‘smiling’). Therefore, it is not surprising that filters learned by attributes lead to effective representations for face localization.



$$h_{i=1\dots 9}^{(1)}$$

*c+

*d+

*e+

Figure 6. Averaged response maps of LNet, including (a) CelebA, (b) MobileFaces, (c) some failure cases.

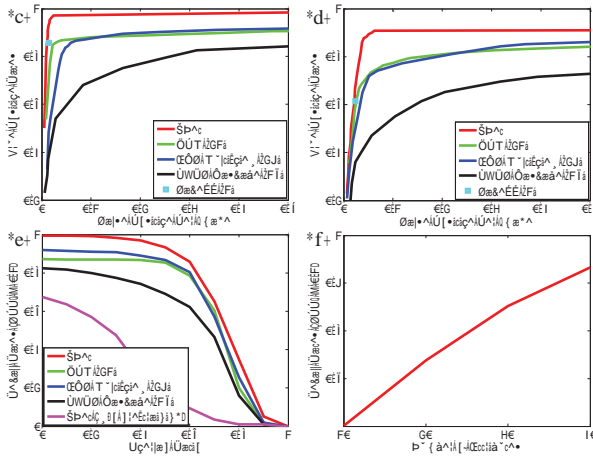


Figure 7. ROC curves on (a) CelebA (b) MobileFaces. (c) Recall rates w.r.t. overlap ratio ($FPP1 = 0.1$). (d) Recall rates w.r.t. number of attributes ($FPP1 = 0.1$)

to the ground truth landmark points. PANDA-w and PANDA-l are based on PANDA [32], which was proposed recently for human attribute recognition by ensembling multiple CNNs, each of which extracts features from a well-aligned human part. These features are concatenated to train SVM for attribute recognition. It is straightforward to adapt this method to face attributes, since face parts can be well-aligned by landmark points. Here, we consider two settings. PANDA-w obtains the face parts by applying the state-of-the-art face detection [17] and alignment [25] on wild images, while PANDA-l attains the face parts by using ground truth landmark points. For fair comparison, all the above methods are trained with the same data as ours.

3.1. Effectiveness of the Framework

This section demonstrates the effectiveness of the framework. All experiments in this section are done on CelebA.

• LNet

Performance Comparison We compare LNet with four state-of-the-art face detectors, including DPM [21], ACF Multi-view [30], SURF Cascade [17], and Face++ [1]. We evaluate them by using ROC curves when $IoU^1 \geq 0.5$. As plotted in Fig.7(a), when $FPP1 = 0.01$, the true

¹IoU indicates Intersection over Union.

positive rates of Face++ and LNet are 85% and 93%; when $FPP1 = 0.1$, our method outperforms the other three methods by 11, 9 and 22 percent respectively. We also investigate how these methods perform with respect to overlap ratio (IoU), following [35, 21]. Fig.7(c) shows that LNet generally provides more accurate face localization, leading to good performance in the subsequent attribute prediction.

Further Analysis LNet significantly outperforms LNet (without pre-training) by 74 percent when the overlap ratio equals to 0.5, which validates the effectiveness of pre-training, as shown in Fig.7(c). We then explore the influence of the number of attributes on localization. Fig.7(d) illustrates rich attribute information facilitates face localization.

To examine the generalization ability of LNet, we collect another 3,876 face images for testing, namely MobileFaces, which comes from a different source² and has a different distribution from CelebA. Several examples of MobileFaces are shown in Fig.6(b) and the corresponding ROC curves are plotted in Fig.7(b). We observe that LNet constantly performs better and still gains 7 percent improvement ($FPP1 = 0.1$) compared with other face detectors. Despite some failure cases due to extreme poses and large occlusions, LNet accurately localize faces in the wild as demonstrated in Fig.6.

• ANet

Pre-training Discovers Semantic Concepts We show that pre-training of ANet can implicitly discover semantic concepts related to face identity. Given a hidden neuron at the FC layer of ANet as shown in Fig.2(c), we partition the face images into three groups, including the face images with high, medium, and low responses at this neuron. The face images of each group are then averaged to obtain the mean face. We visualize these mean faces for several neurons in Fig.8(a). Interestingly, these mean face changes smoothly from high response to low response, following a high-level concept. Human can easily assign each neuron with a semantic concept it measures (i.e. the text in yellow).

²MobileFaces was collected by users with mobile phones, while CelebA and LFWA collected face images of celebrities taken by professional photographers.

Jkij'Tgur0	↔	Nqy'Tgur0	Jkij'Tgur0	↔	Nqy'Tgur0	Vguw'Kocig	Ceikxcvqpu	Pgwtqpu				
*cl3+	Igpfgt	*cl4+	Jckt'Eqsqt	*dl3+				Depiu	Dtqyp'Jckt	Rcng'Unkp	Pcttqy'G{gu	Jkij'Ejggm
*cl5+	Cig	*cl6+	Teeg	*dl4+				G{gincungu	Owucejg	Dccem'Jckt	Uolnkpj	Dki'Pqug
*cl7+	Hceg'Ujerg	*cl8+	G{g'Ujerg	*dl5+				Ygc0'Jcv	Dsqpf'Jckt	Ygc0'Nkrnkem	Culcp	Dki'G{gu

		5 Shadow	Arch. Eyebrows	Attractive	Bags Un. Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	H. Cheekbones	Male
CelebA	FaceTracer [14]	85	76	78	76	89	88	64	74	70	80	81	60	80	86	88	98	93	90	85	84	91
	PANDA-w [32]	82	73	77	71	92	89	61	70	74	81	77	69	76	82	85	94	86	88	84	80	93
	PANDA-I [32]	88	78	81	79	96	92	67	75	85	93	86	77	86	86	88	98	93	94	90	86	97
	[17]+ANet	86	75	79	77	92	94	63	74	77	86	83	74	80	86	90	96	92	93	87	85	95
	LNets+ANet(w/o)	88	74	77	73	95	92	66	75	84	91	80	78	85	86	88	96	92	93	85	84	94
	LNets+ANet	91	79	81	79	98	95	68	78	88	95	84	80	90	91	92	99	95	97	90	87	98
LFWA	FaceTracer [14]	70	67	71	65	77	72	68	73	76	88	73	62	67	67	70	90	69	78	88	77	84
	PANDA-w [32]	64	63	70	63	82	79	64	71	78	87	70	65	63	65	64	84	65	77	86	75	86
	PANDA-I [32]	84	79	81	80	84	73	79	87	94	74	74	79	69	75	89	75	81	93	86	92	
	[17]+ANet	78	66	75	72	86	84	70	73	82	90	75	71	69	68	70	88	68	82	89	79	91
	LNets+ANet(w/o)	81	78	80	79	83	84	72	76	86	94	70	73	79	70	74	92	75	81	91	83	91
	LNets+ANet	84	82	83	83	88	88	75	81	90	97	74	77	82	73	78	95	78	84	95	88	94
		Mouth S. O.	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Reced. Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wear. Earrings	Wear. Hat	Wear. Lipstick	Wear. Necklace	Wear. Necktie	Young		Average
CelebA	FaceTracer [14]	87	91	82	90	64	83	68	76	84	94	89	63	73	73	89	89	68	86	80		81
	PANDA-w [32]	82	83	79	87	62	84	65	82	81	90	89	67	76	72	91	88	67	88	77		79
	PANDA-I [32]	93	93	84	93	65	91	71	85	87	93	92	69	77	78	96	93	67	91	84		85
	[17]+ANet	85	87	83	91	65	89	67	84	85	94	92	70	79	77	93	91	70	90	81		83
	LNets+ANet(w/o)	86	91	77	92	63	87	70	85	87	91	88	69	75	78	96	90	68	86	83		83
	LNets+ANet	92	95	81	95	66	91	72	89	90	96	92	73	80	82	99	93	71	93	87		87
LFWA	FaceTracer [14]	77	83	73	69	66	70	74	63	70	71	78	67	62	88	75	87	81	71	80		74
	PANDA-w [32]	74	77	68	63	64	64	68	61	64	68	77	68	63	85	78	83	79	70	76		71
	PANDA-I [32]	78	87	73	75	72	84	76	84	73	76	89	73	75	92	82	93	86	79	82		81
	[17]+ANet	76	79	74	69	66	68	72	70	71	72	82	72	65	87	82	86	81	72	79		76
	LNets+ANet(w/o)	78	87	77	75	71	81	76	81	72	72	88	71	73	90	84	92	83	76	82		79
	LNets+ANet	82	92	81	79	74	84	80	85	78	77	91	76	76	94	88	95	88	79	86		84

Table 1. Performance comparison of attribute prediction. (Note that FaceTracer and PANDA-I attains the face parts by using ground truth landmark points.)

	Gender	Asian	White	Black	Youth	M. Aged	Senior	Black H.	Blond H.	Bald	No Eye.	Eye.	Mustache	R. Hair.	B. Eye.	A. Eye.	B. Nose	No Beard	R. Jaw		Average
FaceTracer [14]	91	87	86	75	66	54	70	66	68	72	84	86	83	76	72	66	65	81	51		73
POOF [2]	92	90	81	90	71	60	80	67	75	67	87	90	86	72	74	71	68	77	55		76
LNets+ANet	94	85	83	87	80	77	81	86	89	84	85	84	86	83	82	75	79	78	81		83

Table 2. Performance comparison on extended attributes. (Performance are measured by the average of true positive rates and true negative rates.)

uation protocol is the same as [2]. In Table 2, LNets+ANet outperforms them by 10 and 7 percent respectively.

Further Analysis When compared with [17]+ANet, LNets accounts for nearly 6 percentage improvement over using an off-the-shelf face detector [17]. We also experiment with the case of providing ANet with localized face region by LNets, but without pre-training, denoted as LNets+ANet(w/o). The average accuracies have dropped 4 and 5 percent on CelebA and LFWA, which indicate pre-training with massive facial identities helps discover semantic concepts. To further examine whether the proposed approach can be generalized to unseen attributes, we manually label 30 more attributes for the testing images on LFWA. To test on these 30 attributes, we directly transfer weights learned by deep models to extract features, and only re-train SVMs using one third of the images. LNets+ANet leads to 8, 10, and 3 percent average gains over the other three approaches (FaceTracer, PANDA-w, and PANDA-I).

Time Complexity For a 300 × 300 image, LNets takes 35ms to localize face region while ANet takes 14ms to output extracted features on GPU. In contrast, a naïve

patch-by-patch scanning needs nearly 80 ms to extract features. Our framework has large potential in real-world applications.

4. Conclusion

This paper has proposed a novel deep learning framework for **face attribute prediction** in the wild. With carefully designed pre-training strategies, our method is robust to background clutters and face variations. We devise a new fast feed-forward algorithm for locally shared filters to save redundant computation, which enables evaluating image with arbitrary size in realtime. It allows taking images of arbitrary sizes as input without normalization. We have also revealed multiple important facts about learning face representation, which shed a light on new directions of face localization and representation learning.

Acknowledgement This work was partially supported by the National Natural Science Foundation of China (91320101, 61472410, 61503366) and the Research Grants Council of Hong Kong (No. CUHK14207814).

For more technical details, please contact the corresponding author Ping Luo via pluo.lhi@gmail.com.

References

- [1] Face++. <http://www.faceplusplus.com/>.
- [2] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In CVPR, pages 955–962, 2013.
- [3] A. Bergamo, L. Bazzani, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. arXiv preprint arXiv:1409.3964, 2014.
- [4] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In ICCV, pages 1543–1550, 2011.
- [5] J. Chung, D. Lee, Y. Seo, and C. D. Yoo. Deep attribute networks. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, volume 3, 2012.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255, 2009.
- [7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531, 2013.
- [8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. JMLR, 9:1871–1874, 2008.
- [9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In CVPR, pages 1778–1785, 2009.
- [10] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In CVPR, volume 2, pages 1735–1742, 2006.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, pages 346–361. 2014.
- [12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012.
- [14] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In ECCV, pages 340–353. 2008.
- [15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In ICCV, pages 365–372, 2009.
- [16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In NIPS, 1990.
- [17] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In CVPR, pages 3468–3475, 2013.
- [18] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. CVPR, 2012.
- [19] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In ICCV, pages 2864–2871, 2013.
- [20] O. K. Manyam, N. Kumar, P. Belhumeur, and D. Kriegman. Two faces are better than one: Face recognition in group photographs. In IJCB, pages 1–8, 2011.
- [21] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In ECCV, pages 720–735. 2014.
- [22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?—weakly-supervised learning with convolutional neural networks. In CVPR, pages 685–694, 2015.
- [23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. arXiv preprint arXiv:1403.6382, 2014.
- [24] F. Song, X. Tan, and S. Chen. Exploiting relationship between attributes for improved face verification. CVIU, 122:143–154, 2014.
- [25] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In CVPR, pages 3476–3483, 2013.
- [26] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In NIPS, 2014.
- [27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In CVPR, pages 1701–1708, 2014.
- [28] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. CVPR, 2015.
- [29] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, pages 3485–3492, 2010.
- [30] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In IJCB, pages 1–8, 2014.
- [31] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In ECCV, pages 834–849. 2014.
- [32] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In CVPR, 2014.
- [33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In ICLR, 2015.
- [34] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. NIPS, 2014.
- [35] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In ECCV, pages 391–405. 2014.