# LOW DIMENSIONAL DEEP FEATURES FOR FACIAL LANDMARK ALIGNMENT

*Ankit Jalan, Siva Chaitanya Mynepalli, Viswanath Veera, Shankar M. Venkatesan*

Samsung R&D Institute, Bangalore, India
{ankit.jalan, chaitanya.15, viswanath.v, s.venkatesan}@samsung.com

## ABSTRACT

We propose a **L**ow-Dimensional **D**eep **F**eature based **F**ace **A**lignment (LDFFA) method to address the problem of face alignment "in-the-wild". Recently, Deep Bottleneck Features (DBF) has been proposed as an effective channel to represent input with compact, low-dimensional descriptors. The locations of fiducial landmarks of human faces could be effectively represented using low dimensional features due to the large correlation between them. In this paper, we propose a novel deep CNN with a bottleneck layer which learns to extract a low-dimensional representation (DBF) of the fiducial landmarks from images of human faces. We pre-train the CNN with a large dataset of synthetically annotated data so that the extracted DBFs are robust across variations in pose, occlusions, and illumination. Our experiments show that the proposed approach demonstrates near real-time performance and higher accuracy when compared with state-of-the-art results on numerous benchmarks.

***Index Terms***— Face Landmarks, Face Alignment, Face Pose, Convolutional Neural Network (CNN)

## 1. INTRODUCTION

Facial landmark alignment, also called face alignment, serves as an essential preprocessing stage in various tasks such as face recognition, emotion recognition as well as face modelling, beautification, and animation. Consequently it has been extensively studied in the past decade. The problem of face alignment in a constrained environment has been well addressed [27, 38], but, in an unconstrained environment with various confounding factors like pose, occlusion, expression, and illumination, this problem remains challenging.

Inspired by the excellent performance of CNNs in numerous vision related applications, various deep learning based methods have been proposed and shown to be effective in learning to predict the face landmarks [2, 3, 18, 21, 37]. However, the high predictive ability of CNNs is limited by the availability of training data. There are some publicly held hand annotated datasets for face landmarks which are not sufficient to effectively train a CNN which has millions of parameters. While it is a very common practice to perform data augmentation with translation, in-plane rotation etc., the resulting images lack variation in face shape and texture and thus the network trained on this data might not be able to generalize under some conditions.

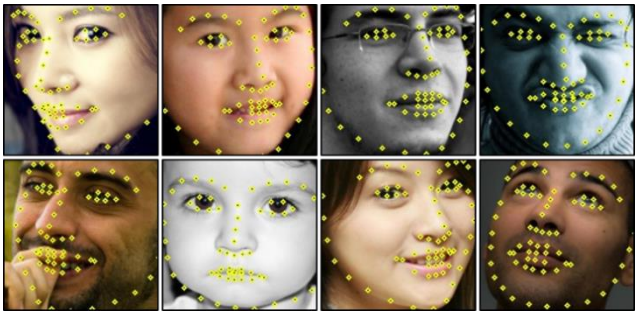


Fig. 1. Selected results **on 300-W (full-set). Our LDFFA** method is used to detect 68 landmarks

In this work, we propose a Low-Dimensional Deep Feature based Face Alignment (LDFFA) method to address the problem of face alignment "in-the-wild". The CNN architecture used in this method extracts low-dimensional deep bottleneck features (DBF) to estimate the actual landmark locations. We pre-train the network with a large database of synthetically annotated images generated using a regression based algorithm to fully harness the CNN's predictive ability. We then fine-tune the network with the hand annotated datasets.

The DBF extracted from this network are shown to have improved representation power [39] and are robust to variations in the environment. Further, LDFFA is independent of shape initialization. These factors ensure robust performance of LDFFA in case of large pose and illumination variation.

Experimental analysis has demonstrated that LDFFA outperforms other state-of-the-art algorithms on Helen, LFPW benchmark datasets while giving comparable performance on subsets of 300-W database [26, 27, 28, 31] with 68 fiducial landmarks. In the following section, we analyze various related works. In Section 3, we formally explain LDFFA, and discuss dataset generation. Experimental analysis and comparison with other state-of-the-art methods is presented in Section IV. Finally, in Section V, we conclude the paper.

## 2. RELATED WORK

Numerous methods have been proposed to tackle the problem of face alignment with varying degrees of success. Overall, face alignment can be formulated as a problem of searching pre-defined landmarks in a face image. There are several methods proposed to solve this problem and they can be broadly classified as **discriminative methods** such as *Constrained Local Models (CLM)* based, *Regression* based, or *Deep CNN* based methods and **generative methods** such as *Active Appearance Model (AAM)* based methods.

**CLM** based methods [4, 5] learn independent local detectors for each facial points and they regularize the detection responses of each local detector using a parametric (PCA based) shape model [5] or an exemplar based model [27]. **AAM** based methods [6, 7, 8], initially proposed by Cootes *et al*. [6], employ linear statistical models of both shape and appearance of deformable object. They are widely used in Computer vision tasks as they are able to generate a wide variety of instances using a few model parameters. Face alignment by **Regression** based methods has seen significant progress in recent years due to the availability of large datasets with great variation in face poses. Most of these methods employ a cascaded regression strategy as it is shown to generalize well and they are also time efficient. There are many methods that belong to this family [9, 10, 11, 12, 13, 14, 15, 16, 17]; however a classic work in this area is Supervised Descent Method (SDM) by Xiong *et al.* [15], which was the first work to describe the cascaded regression problem as a general framework for optimizing non-linear objective functions. In this work, regressors at each level of the cascade are assumed to be linear and they model the average descent directions. SDM uses local SIFT features extracted around the current estimate of the shape to predict an update to it. Global SDM [16] improves on the SDM by dividing search space into regions of similar gradient directions. Instead of using SIFT features for regression, Ren *et al.* [14] proposed to learn the local binary features with random forests, which resulted in improved computation time with greater accuracy. These methods generally use a mean-shape initialization which could result in poor performance in case of extreme pose when the actual shape is very different from the mean-shape. To circumvent this problem, Cao *et al.* [11] proposed an algorithm with different initializations and takes the median of all the predictions as final output. Zhu *et al*. [17] proposed a coarse-to-fine shape search which finds the best possible initialization at each level.

Several methods based on Deep Learning have been proposed for face alignment. Sun *et al*. [2] proposed a three-level cascaded Deep CNN in which the first level gives an initialization and subsequent levels work on local patches around the initial estimate to further refine it. Kumar *et al.* [18] recently proposed an algorithm in which they use Deep CNN to extract features from local patches, which could be used to replace SIFT features. Lai *et al*. [24] used a similar idea to extract local features from the current estimate of the shape. However, in their approach the local features were sampled directly from the output of one of the de-convolution layers of the CNN which predicts the initial estimate, which allows them to circumvent the problem of initialization.

In the following sections, we show that our method achieves comparable results to most of these based works without the added complexity of a cascade of regressors.

## 3. LDFFA

From our experiments we observed that a set of **k** (> 50) basis vectors, obtained by performing a Principal Component Analysis (PCA) on the coordinates of fiducial landmarks of faces, were sufficient to represent the data to within an error of 1% of the inter-ocular distance (using equation 1 in Section 4.2). Based on this observation, we deduce that the coordinates of the facial landmarks are highly correlated and could be efficiently described with a low-dimensional representation. This served as the main motivation to employ a CNN to extract the coefficients of the eigen vectors from images of faces, which could then be used to predict the landmark coordinates. However, we realized that a network that is trained end-to-end would predict the locations of landmarks with better accuracy. Consequently, we employed a bottle-neck architecture for the CNN to extract Deep Bottleneck Features which were mapped to landmark coordinates.

### 3.1. Bottleneck Architecture

The network used in LDFFA method consists of eight convolutional layers with four max-pooling layers placed in-between them. Features extracted by the CNN are then fed into a network of three fully connected (FC) layers and an output layer (Fig. 2). In this architecture, the convolutional layers can be regarded as global feature extractors for the face image, and the bottleneck layer can be considered as an encoder. The narrow shape of the bottleneck layer ensures that the network learns the low dimensional deep features which could describe the face shape.

### 3.2. Implementation Details

As shown in Fig. 2, our network topology consists of four sets of *Conv-LReLU-Conv-LReLU-Pool* layer. These four sets employ 32, 64, 128 & 256 filters respectively. Each set has two convolutional layers with kernel size of 3×3, a LeakyReLU (LReLU) layer ($\alpha = 0.01$) and a 2×2 max-pool layer with stride 2. We use 3 fully connected layers with 200, 200 & 50 neurons respectively which are then connected to the output layer of size 136. Except for the last fully connected layer which employs linear activation, all the activations are LReLU ($\alpha = 0.01$). Input to the LDFFA-

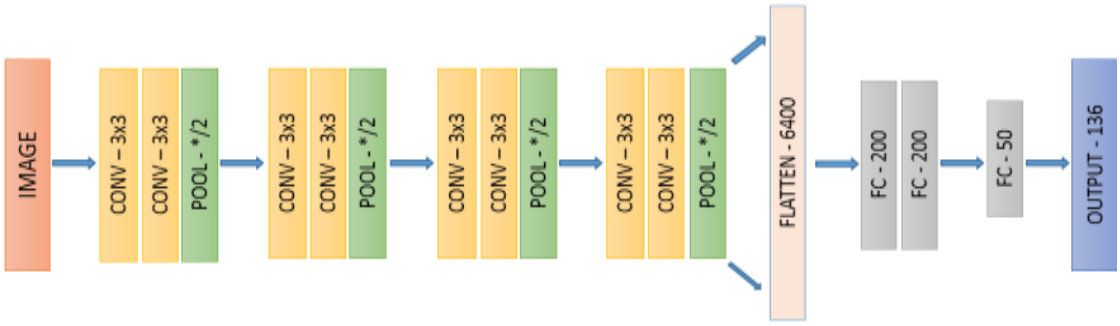



**Fig. 2.** Overview of the proposed deep convolutional neural network architecture (LDFFA) for face alignment. The network takes an image input and directly estimates the coordinates of facial landmarks.

network is an image of size 128×128 which is generated by cropping and resizing the bounding box of the face. The output is a 136 dimensional face landmark location in x-y format.

To learn the weights, we used stochastic optimization with *Adam* [29] optimizer provided by [30] with default parameters. Adam had better performance over Stochastic Gradient Descent (SGD) and *Adadelta* on the validation set. Also a per-pixel Gaussian noise of $\sigma = 0.5$ was added to the input images to further augment the training images.

To enhance the predictive ability of the CNN, we pre-trained it on 164K and validated on about 17K randomly selected images from the IMDB [22, 23] dataset. An off-the-shelf face detector [25] was used to identify the bounding box (which was enlarged to 1.2x) of faces in these images and the facial landmarks were detected using Kazemi *et al.*'s [12] algorithm's implementation provided by D-lib [25]. The CNN trained in this manner is expected to be robust because of the large variation in pose, texture, and illumination in the database. The network weights were then fine-tuned with the training set of the 300-W database as explained in Section 4.1. The bounding boxes for 300-W images were provided by [28, 31] using their in-house face detector.

## 4. EXPERIMENTS

In order to evaluate the performance of LDFFA, we perform rigorous experiments and compare it with the results of various state-of-the-art methods. The evaluations were done on the three widely used benchmark datasets. These datasets have large variations in illumination, occlusions & head pose.

### 4.1. Datasets

Pre-training of LDFFA network was done on 164K randomly picked images from the IMDb database [22, 23].

**Helen** dataset [26]: 2000 training and 330 testing images with variations in pose and illumination.

**LFPW** dataset [27]: 811 training and 224 testing images provided by [28].

**300-W** dataset [28, 31]: This dataset is created from existing dataset (LFPW, Helen, AFW, and XM2VTS) and a new dataset called IBUG.

For evaluation and a fair comparison with other methods we follow the same data configuration as in [17]. Our training set consists of AFW, training set of LFPW and training set of Helen database with a total of **3148** images. Our testing set consists of testing set from Helen, testing set have large variations in illumination, occlusions & head pose from LFPW and IBUG dataset with a total of **689** images.

### 4.2. Evaluation

We follow the method of [31] where the average L2 distance of the estimated landmark position from the ground truth is normalized by the standard definition of inter-ocular distance ($d_{outer,}$) to give the error (Eqn. 1). For each of the benchmark dataset, we report the mean error evaluated over all the images. Also, to compare the results with papers reporting cumulative error distribution (CED) performance, we plot CED curves for the subset of 300W test dataset

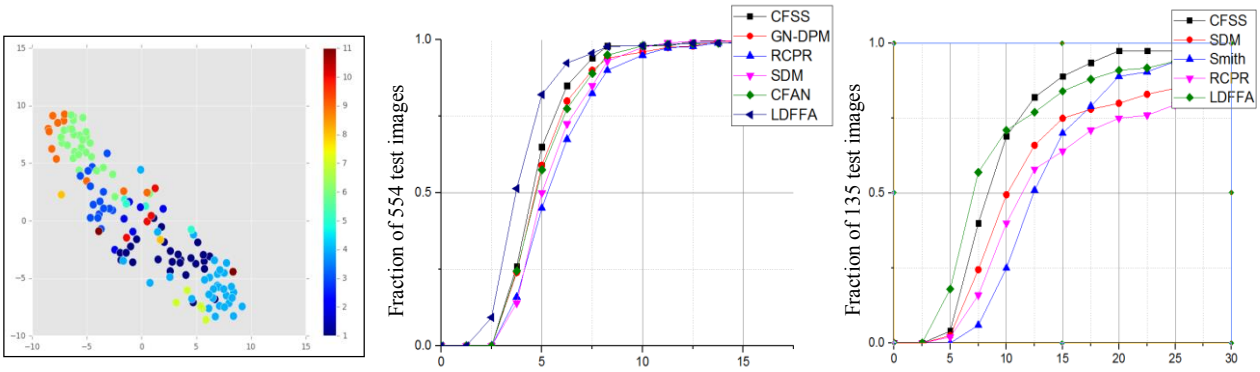
$$E = \sum_{i=1}^{N_i} \frac{\| X_p - X_g \|_2}{d_{outer} N} \quad (1)$$

Where, $\boldsymbol{X_p}$ is the predicted landmark location and $\boldsymbol{X_g}$ is the corresponding ground truth value. $N$ is the number of facial landmarks, here 68. $d_{outer}$ is the L2 distance between the outer eye corners.

### 4.3. Comparison

We evaluate LDFFA against the performance of other state-of-the-art face alignment methods.

Some of the deep CNN based methods [1, 2, 3] mainly detect 5 facial landmarks and hence the results are not comparable. We compare our results with those reported in




(a) t-SNE depiction of internal states. (Challenging subset, 300-W)

(b) CED for 68-pts common subset of 300W

(c) CED for 68-pts challenging subset of 300W

**Fig. 3.** (a) t-SNE depiction of internal states shows clustering of the images (Best viewed in color), (b & c) Comparison of cumulative error distribution curves. Proposed LDFFA method has higher accuracy than state-of-the-art methods.

[17, 18]. Table 1 and Fig. 3 (b & c) illustrate the Normalized Mean Error using definition from [31].

Please note that the data for normalized mean error for Table 1 & CED graph has been obtained from published literature [17] and to make a fair comparison we sampled the data from their graphs to plot the CED curves in Fig.3. It is evident from the CED curve that our LDFFA method has a higher average accuracy than many of the above mentioned state-of-the-art methods.

**Table 1.** The normalized mean error on LFPW, Helen and 300-W dataset (*Co.: Common, Ch.: Challenging, Fu: Full Set*) with First and Second best results highlighted

| Methods | LFPW | HELEN | Co. | Ch. | Fu. |
|---|---|---|---|---|---|
| Zhu et. al. [32] | 8.29 | 8.16 | 8.22 | 18.33 | 10.20 |
| DRMF [33] | 6.57 | 6.7 | 6.65 | 19.79 | 9.22 |
| ESR [11] | -- | -- | 5.28 | 17.00 | 7.58 |
| RCPR [10] | 6.56 | 5.93 | 6.18 | 17.26 | 8.35 |
| SDM [15] | 5.67 | 5.50 | 5.57 | 15.40 | 7.50 |
| Smith et al [34] | -- | -- | -- | 13.30 | -- |
| Zhao et al [35] | -- | -- | -- | -- | 6.31 |
| GN-DPM [36] | 5.92 | 5.69 | 5.78 | -- | -- |
| CFAN [19] | 5.44 | 5.53 | 5.50 | -- | -- |
| ERT [12] | -- | -- | -- | -- | 6.40 |
| LBF [14] | -- | -- | 4.95 | 11.98 | 6.32 |
| CFSS [17] | 4.87 | 4.63 | 4.73 | **9.98** | 5.76 |
| LDDR [18] | 4.67 | 4.76 | -- | 11.49 | -- |
| **LDFFA** | **4.24** | **4.01** | **4.10** | **9.99** | **5.26** |

### 4.5. Runtime

All the experiments were performed using an NVIDIA TESLA-K80 GPU. We used *Keras: Deep learning library* [30] to implement the above network. Pre-training on 164K images and fine-tuning on 3148 images from 300W took about 8 hours. During the testing phase, in GPU, LDFFA takes around 1.2 milliseconds and in CPU (Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz), LDFFA takes around 40 milliseconds achieving a near real-time performance.

## 5. DISCUSSION & CONCLUSION

We propose a deep CNN based architecture with unique bottle-neck feature for face alignment problem. Database generation for pre-training is facilitated by employing existing algorithms. From Table 1 we can observe that our method outperforms other methods in case of LFPW, Helen and Common set of 300W database, in terms of accuracy. While LDFFA performs second best in case of 300W challenging subset, the average accuracies are comparable.

We used t-SNE plot to analyze the deep features extracted from an image by the convolution layers of CNN, (Fig. 3(a)). Each image is labeled by performing K-means clustering of the landmark points. The colors in the t-SNE plot correspond to the label provided by the K-means clustering. It can be observed from the Fig. 3(a) that our deep features are able to differentiate and cluster images with similar landmarks distribution together. This shows that these features encode the possible landmark distributions.

From the CED plot for the challenging dataset, it can be observed that the proposed method is outperformed in extremely challenging cases, where other methods have a higher fraction of data at relatively lower errors. One possible solution to this issue could be to refine the estimate of LDFFA using local features, like methods based on cascade of regressors.

## 6. REFERENCES


[1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. "Facial landmark detection by deep multi-task learning." *ECCV*, Springer International, pp. 94-108, 2014.

[2] Y. Sun, X. Wang, and X. Tang. "Deep convolutional network cascade for facial point detection." IEEE *CVPR*, pp. 3476-3483. 2013.

[3] Z. Shao, S. Ding, H. Zhu, C. Wang, and L. Ma. "Face alignment by deep convolutional network with adaptive learning rate." IEEE *ICASSP*, pp. 1283-1287, 2016.

[4] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models." Pattern Recognition 41, no. 10 (2008): 3054-3067.

[5] J. M. Saragih, S. Lucey, and J. F. Cohn. "Deformable model fitting by regularized landmark mean-shift." International Journal of Computer Vision 91, no. 2 (2011): 200-215.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. "Active appearance models." IEEE Trans. on *PAMI* 23, no. 6 (2001): 681-685.

[7] E. Antonakos, J. A. Medina, G. T., and S. P. Zafeiriou, "Feature-based lucas-kanade and active appearance models" IEEE Trans. on Image Processing 24, no. 9 (2015): 2617-2632.

[8] I. Matthews and S. Baker. "Active appearance models revisited". International Journal of Computer Vision 60, no. 2 (2004): 135-164.

[9] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. "Incremental face alignment in the wild." IEEE *CVPR*, pp. 1859-1866. 2014.

[10] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. "Robust face landmark estimation under occlusion." IEEE *ICCV*, pp. 1513-1520. 2013.

[11] X. Cao, Y. Wei, F. Wen, and J. Sun. "Face alignment by explicit shape regression." International Journal of Computer Vision 107, no. 2 (2014): 177-190.

[12] V. Kazemi and J. Sullivan. "One millisecond face alignment with an ensemble of regression trees." IEEE *CVPR*, pp. 1867-1874. 2014.

[13] D. Lee, H. Park, and C. D. Yoo. "Face alignment using cascade gaussian process regression trees." IEEE *CVPR*, pp. 4204-4212. 2015.

[14] S. Ren, X. Cao, Y. Wei, and J. Sun. "Face alignment at 3000 fps via regressing local binary features." IEEE *CVPR*, pp. 1685-1692. 2014.

[15] X. Xiong and F. De la Torre. "Supervised descent method and its applications to face alignment." IEEE *CVPR*, pp. 532-539. 2013.

[16] X. Xiong, and F. De la Torre. "Global supervised descent method." IEEE *CVPR*, pp. 2664-2673. 2015.

[17] S. Zhu, C. Li, C. C. Loy, and X. Tang. "Face alignment by coarse-to-fine shape searching." IEEE *CVPR*, pp. 4998-5006. 2015.

[18] A Kumar, R Ranjan, V Patel, & R Chellappa. "Face Alignment by Local Deep Descriptor Regression.", arXiv: 1601.07950 (2016).

[19] J. Zhang, S. Shan, M. Kan, and X. Chen. "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment." In ECCV, pp. 1-16. Springer International, 2014.

[20] Itseez, "OpenCV" https://github.com/itseez/opencv

[21] Y Wu, Z Wang, and Q Ji. "Facial feature tracking under varying facial expressions and face poses based on restricted Boltzmann machines." IEEE *CVPR*, pp. 3452-3459. 2013.

[22] R. Rothe, R. Timofte, and L. V Gool. "DEX: Deep EXpectation of apparent age from a single image." IEEE ICCV, pp. 10-15. 2015.

[23] R. Rothe, R. Timofte, and L. V Gool. "Deep Expectation of Real and Apparent Age from a Single Image without Facial Landmarks." IJCV (2016): 1-14.

[24] H Lai, S Xiao, Z Cui, Y Pan, C Xu, and S Yan. "Deep Cascaded Regression for Face Alignment.", arXiv: 1510.09083 (2015).

[25] King, Davis E. "Dlib-ml: A machine learning toolkit." Journal of Machine Learning Research 10, no. Jul (2009): 1755-1758.

[26] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. "Interactive facial feature localization." In ECCV, pp. 679-692. Springer Berlin Heidelberg, 2012.

[27] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. "Localizing parts of faces using a consensus of exemplars." IEEE *PAMI* 35, no. 12 (2013): 2930-2940.

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. "300 faces in-the-wild challenge: The first facial landmark localization challenge." IEEE *ICCV*, pp. 397-403. 2013.

[29] D. Kingma, and J. Ba. "Adam: A method for stochastic optimization." arXiv: 1412.6980 (2014).

[30] F. Chollet, "Keras.", https://github.com/fchollet/keras, 2015

[31] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. "300 faces in-the-wild challenge: Database and results." Image and Vision Computing 47 (2016): 3-18.

[32] X. Zhu, and D Ramanan. "Face detection, pose estimation, and landmark localization in the wild." IEEE *CVPR*, pp. 2879-2886, 2012.

[33] A. Asthana, S. Zafeiriou, S. Cheng, & M. Pantic, "Robust discriminative response map fitting with constrained local models." IEEE *CVPR*, pp. 3444-3451. 2013.

[34] B.M. Smith, J. Brandt, Z. Lin, and L. Zhang. "Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization." IEEE *CVPR*, pp. 1741-1748. 2014.

[35] X Zhao, T Kim, & W Luo "Unified face analysis by iterative multi output random forests." IEEE *CVPR*, pp. 1765-1772. 2014.

[36] G. Tzimiropoulos and M. Pantic. "Gauss-newton deformable part models for face alignment in-the-wild." IEEE *CVPR*, pp. 1851-1858. 2014.

[37] J W. Baddar, J Son, D H Kim, S T Kim, and Y M Ro. "A deep facial landmarks detection with facial contour and facial components constraint." IEEE *ICIP*, pp. 3209-3213, 2016.

[38] X. Jin and X. Tan. "Face Alignment In-the-Wild: A Survey." arXiv: 1608.04188 (2016).

[39] Y. Song, I. McloughLin, and L. Dai. "Deep Bottleneck Feature for Image Classification." In Proc. of the 5th ACM on ICMR, pp. 491-494. ACM, 2015.