

Two-Stream Deep Correlation Network for Frontal Face Recovery

Ting Zhang, Qulei Dong, Ming Tang, and Zhanyi Hu

Abstract—Pose and textural variations are two dominant factors to affect the performance of face recognition. It is widely believed that generating the corresponding frontal face from a face image of an arbitrary pose is an effective step toward improving the recognition performance. In the literature, however, the frontal face is generally recovered by only exploring textural characteristic. In this letter, we propose a two-stream deep correlation network, which incorporates both geometric and textural features for frontal face recovery. Given a face image under an arbitrary pose as input, geometric and textural characteristics are first extracted from two separate streams. The extracted characteristics are then fused through the proposed multiplicative patch correlation layer. These two steps are integrated into one network for end-to-end training and prediction, which is demonstrated effective compared with state-of-the-art methods on the benchmark datasets.

Index Terms—Correlation layer, deep neural network, frontal face recovery, geometric stream, textural stream.

I. INTRODUCTION

FACE recognition is a field of great potential, which has been widely used in access control, video surveillance, personal verification, etc. Over the past decade, there have been tremendous advances in face recognition, most of which are owed to the development of deep learning [1]–[5]. Although data-driven features extracted by deep neural networks show great advantages over the hand crafted ones in face recognition [6]–[10], the performance of face recognition is usually influenced by the large variations in pose, illumination, expression, etc. Among them, pose variation has been a persistent challenge because it

Manuscript received May 14, 2017; revised July 10, 2017; accepted July 20, 2017. Date of publication August 7, 2017; date of current version August 29, 2017. This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02070002, and in part by the National Natural Science Foundation of China under Grant 61421004, Grant 61375042, and Grant 61573359. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sumohana S. Channappayya. (*Corresponding author: Qulei Dong.*)

T. Zhang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ting.zhang@nlpr.ia.ac.cn).

Q. Dong and Z. Hu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qldong@nlpr.ia.ac.cn; huzy@nlpr.ia.ac.cn).

M. Tang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tangm@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this letter are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/LSP.2017.2736542

may make the intraperson variance exceed the interperson one. In view of this, many methods have been proposed to transfer a face image of arbitrary pose to the frontal one. These methods can be roughly classified into two groups: Two-dimensional (2-D)-based methods [11]–[17] and 3-D-based ones [18]–[21].

The 2-D-based techniques usually encode a test image with some exemplars, or use 2-D image matching algorithms to address the pose variation. In [12], Markov random fields was applied to infer the frontal face images. Li *et al.* [15] proposed an elastic matching method which aligned the patches and matched the face images of different poses based on Gaussian Mixture Model. In [1], a deep convolutional neural network was proposed to recover the frontal image of neutral illumination from those with arbitrary poses and illumination. In [17], recurrent neural networks were combined with autoencoders to render sequences of rotated face images through incremental 3-D rotations.

The 3-D-based techniques attempt to match the captured 3-D facial data to probe face images or align a probe face image to a 3-D face model. Asthana *et al.* [19] constructed an aligned 3-D face model from a nonfrontal face image, and then rotated the model to render a frontal face image. In [20], a virtual view for the probe image was generated based on a set of 3-D displacement fields sampled from a 3-D face database and the synthesized faces were tested.

Despite the demonstrated success, the performance of existing methods on frontal face recovery is still limited. The methods based on 3-D reconstruction are time consuming and sometimes require several views captured at multiple poses. Although being efficient and only requiring a single input image, the performance of 2-D reconstruction methods is limited because they only exploit the facial textures to align face images. These textures are not effective enough to locate correspondence when face is under out-of-plane rotation.

In this letter, we propose a two-stream deep correlation network (TSDCN) to solve the aforementioned limitations. Given an input face image, we extract the textural and geometric features independently via two streams. The textural stream performs similarly with existing methods and the geometric stream predicts the angles of the face poses. The angle predictions are then correlated with the texture correspondence to predict the recovered face image. Experimental results on the Multi-PIE and labeled faces in the wild (LFW) datasets demonstrate the validity of the proposed method.

The contributions of this work include the following.

- 1) We propose a two-stream network to tackle the frontal face recovery problem, which could independently capture textural and geometric features of input face image.

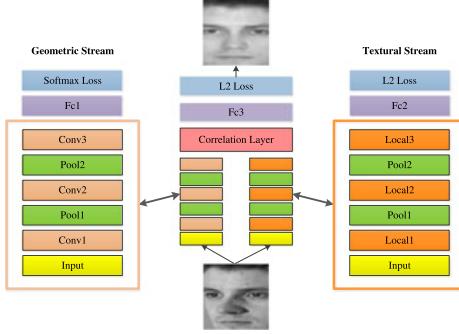


Fig. 1. Network architecture of proposed TSDCN. A face image under arbitrary pose is used as input. After passing the face image through two streams (geometric stream and textural stream), the intermediate feature maps of two streams are correlated to generate the frontal face image. The network parameters can thus be updated by jointly optimizing the ℓ_2 -losses and softmax loss.

- 2) We utilize multiplicative patch correlation to fuse geometric and textural streams for end-to-end face recovery.
- 3) Experiments on the benchmarks demonstrate the superior performance of our method compared with state-of-the-art methods.

Our two-stream scheme seems to have some neurophysiological basis. In [22], it is shown that the response of a face selective neuron in IT can be adequately expressed as a dot product of its average triggered spikes to the 50-D input face vector, whose first 25 elements are of face geometric information, and last 25 ones are of appearance information.

II. PROPOSED ALGORITHM

As shown in Fig. 1, our model TSDCN consists of two streams (geometric stream and textural stream). An input face image is processed at first, and geometric and textural features are extracted separately by two streams. Then, the intermediate layers of the two streams are combined using a correlation layer [23] to reconstruct the frontal face image. Finally, the linear discriminant analysis (LDA) method is used to classify the output image of TSDCN into a certain identity. Our TSDCN will be elaborated in the following sections.

Note that the geometric and textural streams predict in parallel when generating the final face images. Such a design allows us to take the geometric transformation among different poses into account. With the proposed architecture, we constrain the network to extract geometric and textural information separately at first, and then correlate them at a higher stage to generate the recovered face image.

A. Geometric and Textural Streams

The proposed architecture takes 60×60 grayscale face images as inputs. The geometric stream learns to predict the pose of a face image. This stream is composed of three convolutional layers and a fully connected layer. Each convolutional layer is followed by a PReLU activation function [24]. The third convolutional layer is connected to a fully connected layer which contains seven nodes each of which represents the probability of a certain range of possible poses.

In the training phase, the textural stream learns to generate a frontal image from the input image under an arbitrary pose. The input of this stream is the same as that of the geometric stream. This stream is composed of three locally connected layers and a fully connected layer. The features extracted by

TABLE I
ARCHITECTURE OF PROPOSED TSDCN

	Textural stream	Geometric stream
LC	7×7 local, chans 32	Convolution 11×11 conv, chans 32
Pooling	3×3 max pool, stride 2	Pooling 3×3 max pool, stride 2
LC	5×5 local, chans 32	Convolution 11×11 conv, chans 32
Pooling	3×3 max pool, stride 2	Pooling 3×3 max pool, stride 2
LC	5×5 local, chans 32	Convolution 11×11 conv, chans 32
Correlation		5×5 kernel
FC		3600D
Reshape		60×60

The “LC” and “FC” denote the locally connected layer and fully connected layer.

locally connected layers are more powerful than those extracted by standard convolutional layers used in [25]–[27]. More details of the architecture are shown in Table I.

B. Feature Map Fusion

We then fuse the feature maps taken from the last convolutional layer in the geometric stream and the last locally connected layer in the textural stream to enrich the representation. The fused representation captures multiplicative patch correlations between different representations. We denote the geometric and textural representations as Ψ_g and Ψ_t respectively, the correlation between Ψ_g and Ψ_t in a $(2k+1) \times (2k+1)$ patch is defined as

$$\begin{aligned} \Psi_c(x_1, x_2) = & c(\Psi_g, \Psi_t) \\ = & \sum_{o \in [-k, k] \times [-k, k]} \langle \Psi_g(x_1 + o) \Psi_t(x_2 + o) \rangle. \end{aligned} \quad (1)$$

To reduce the computational cost in the correlation, the maximum displacement for comparisons is limited, and the stride is used in both feature maps.

A fully connected layer is applied on Ψ_c as

$$Y = f(W_1 \Psi_c + b_1) \quad (2)$$

where Y is the output of the fully connected layer, $f(\cdot)$ denotes the nonlinear activation function, W_1 the weight matrix, and b_1 the bias vector.

The cost function is defined as

$$L_f = \sum_{j=1}^N \|Y_{j,GT} - Y_j\|^2 \quad (3)$$

$$L = L_f + \alpha L_g + \beta L_t \quad (4)$$

where $Y_{j,GT}$ and Y_j are the ground-truth and generated frontal face image, j and N are the index of the training input and batch size respectively. L_f denotes the loss of the final output image, L_g and L_t the loss of geometric stream and textural stream, respectively. L is the weighted sum of different losses. α and β are constant coefficients to balance the loss function.

III. EXPERIMENTS

We implement our experiments using Caffe [32] and train the proposed TSDCN in three procedures. First, we only enable the supervision from L_g to pretrain the geometric stream. Second, we only exploit the supervision from L_t to pretrain the textural stream. Finally, we fine-tune the whole network with

TABLE II
FACE RECOGNITION RATES (%) UNDER DIFFERENT POSES ON THE
MULTI-PIE DATASET

	-45°	-30°	-15°	$+15^\circ$	$+30^\circ$	$+45^\circ$	Avg
Li [28]	63.5	69.3	79.7	75.6	71.6	54.6	69.3
Zhu [1]	67.1	74.6	86.1	83.3	75.3	61.8	74.7
CPI [11]	66.6	78.0	87.3	85.5	75.8	62.3	75.9
CPF [11]	73.0	81.7	89.4	89.5	80.4	70.3	80.7
VS2VI [13]	62.3	84.3	92.3	91.1	80.5	58.4	78.1
GEM [29]	63.8	75.1	84.7	84.9	75.2	63.0	74.4
HPN [30]	71.3	78.8	82.2	86.2	77.8	74.3	78.4
PPDN [31]	72.1	85.4	92.4	91.4	87.1	71.0	83.2
Textural stream	58.2	80.5	87.3	90.9	78.9	56.0	75.3
TSDCN	69.2	88.9	93.3	95.5	89.0	69.6	84.3

TABLE III
FACE RECOGNITION RATES (%) ON DIFFERENT ILLUMINATIONS ON THE
MULTI-PIE DATASET

	00	01	02	03	04	05	06
Li [28]	51.5	49.2	55.7	62.7	79.5	88.3	97.5
Zhu [1]	72.8	75.8	75.8	75.7	75.7	75.7	75.7
CPI [11]	66.0	62.6	69.6	73.0	79.1	84.5	86.6
CPF [11]	59.7	70.6	76.3	79.1	85.1	89.4	91.3
Textural stream	57.3	64.0	69.9	75.2	79.4	83.2	85.6
TSDCN	69.1	76.2	80.5	83.7	86.1	90.7	91.9
	08	09	10	11	12	13	14
Li [28]	97.7	91.0	79.0	64.8	54.3	47.7	67.3
Zhu [1]	75.7	75.7	75.7	75.7	75.7	75.7	73.4
CPI [11]	86.5	84.2	80.2	76.0	70.8	65.7	76.1
CPF [11]	92.3	90.6	86.5	81.2	77.5	72.8	82.3
Textural stream	86.2	84.7	80.8	75.8	71.0	66.2	77.7
TSDCN	92.1	91.4	89.8	86.6	82.6	76.0	85.1
	15	16	17	18	19		Avg
Li [28]	67.7	75.5	69.5	67.3	50.8		69.3
Zhu [1]	73.4	73.4	73.4	72.9	72.9		74.7
CPI [11]	78.2	80.7	79.4	77.3	65.4		75.9
CPF [11]	84.2	86.5	85.9	82.9	59.2		80.7
Textural stream	78.7	79.8	80.0	78.1	57.5		75.3
TSDCN	85.9	88.5	88.0	88.8	68.9		84.3

the supervision from L . This training manner helps preserve the geometric and textural information for the two streams respectively, and thus enhance the performance. We evaluate the effectiveness of our TSDCN on the popular constrained dataset Multi-PIE [33] and the unconstrained dataset LFW [34].

A. Multi-PIE

The popular Multi-PIE dataset [33] contains images of 337 people with different poses, illuminations, and expressions. Similar to [1] and [11], all the referred methods are evaluated on a subset of Multi-PIE, the same as Setting 1 in [11]: only images under all the 7 poses and 20 illuminations with the neutral expression in session one are adopted for training and testing. In this section, We evaluate the proposed algorithm against the state-of-the-art methods and visualize the features extracted from different layers.

1) *Face Recognition*: In the test stage, features are extracted from the output layer to recover the frontal face images. LDA is used to classify the generated image into a particular identity. Table II reported the recognition rates of the mentioned methods for different poses and Table III for various illuminations. Best results are written in bold. The TSDCN model performs

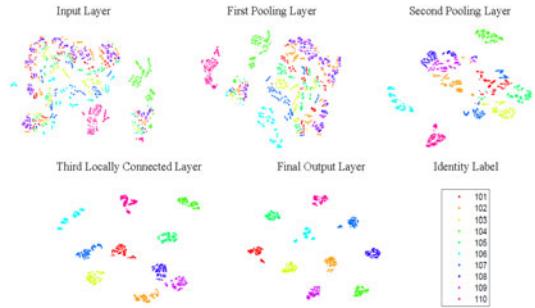


Fig. 2. Feature map visualizations in the textural stream. The dot of the same color represents the same identity. For example, the red dot represents the feature extracted from the images of identity with ID 101.

favorably against the state-of-the-art methods and original textural stream on the Multi-PIE dataset.

As shown in Table II, TSDCN outperforms the state-of-the-art methods [1], [11] on four poses ($+15^\circ$, $+30^\circ$, -15° , -30°), and the average recognition rate. We note that our performance on pose -45° and 45° ranks in the second place. We can conclude from Table II that TSDCN improves the face recognition rates from 3.9% to 8.6% on four poses. Also, the average face recognition rate is increased by 3.6%. TSDCN also outperforms the original textural stream on all the poses, and it indicates the effectiveness of introduction of geometric information.

As shown in Table III, TSDCN outperforms the methods [1], [11] for 15 out of 19 illuminations. Note that the neutral illumination (ID 07) is not included in the test dataset. The results demonstrate the effectiveness of our proposed method.

To investigate the quality of feature representation for face recognition, we use the t-SNE algorithm [35] to transform the feature map from each layer in the textural stream and also the final recovered frontal images from high-dimensional space into 2-D space, so that similar samples are spatially clustered. As shown in Fig. 2, the mixed features from the first pooling layer are similar to those from the input layer. Features of the same identity begin to merge with each other from the second pooling layer. In the third locally connected layer, features are basically separated from each other. The reconstructed frontal images predicted from the same identity clustered, while those from different identities are spatially distinct. The above results show that the textural stream can extract effective textural information.

2) *View Classification*: The TSDCN's ability to predict the poses from the input images is evaluated on the Multi-PIE dataset. Table V reported the classification accuracies of different poses by TSDCN. The classification rates of most poses by TSDCN are above 90%. It suggests that we can accurately learn pose angles through the geometric stream.

To investigate the quality of feature representation for view classification, we use the t-SNE method to visualize features extracted from each layer in the geometric stream, and demonstrate that geometric stream can extract effective geometric information. As shown in Fig. 3, features under the same pose gradually merge with each other as the layer increases. The visualization results of the output layer in the geometric stream show that these features are discriminative among different poses.

3) *Evaluation on the Recovered Frontal Images*: Fig. 4 shows several reconstructed examples. It indicates that TSDCN can preserve each subject's characteristics and make the recovery under neutral illumination. To evaluate the quality of the recovered images, LDA is applied on the original face images and the recovered ones. The recognition rates reported in Fig. 5

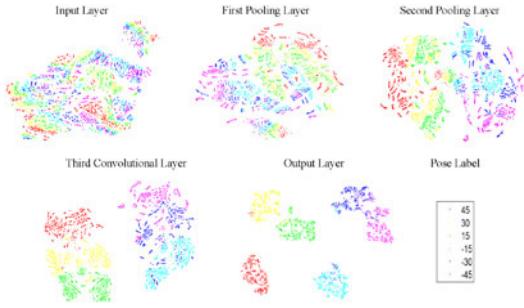


Fig. 3. Feature map visualizations in the geometric stream. The dot of the same color represents the feature of input images under the same viewing condition. For example, the red dot represents the feature extracted from the images under pose 45° .



Fig. 4. Reconstructed examples on the Multi-PIE dataset: (top) Images of six poses under arbitrary illuminations for each identity. (bottom) Reconstructed frontal face images under neutral illumination.

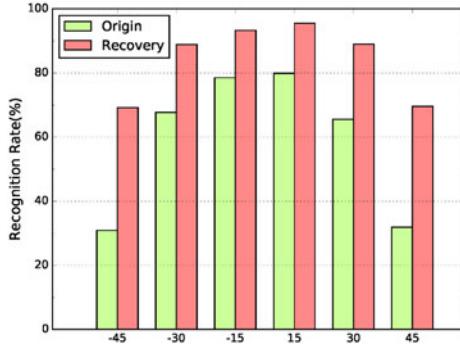


Fig. 5. Face recognition performance on the Multi-PIE dataset by applying LDA on the original face images and recovered ones.

demonstrate the advantage of our recovered images over the original images. It indicates that frontal recovered face images improve the recognition performance more when the input face images are in larger pose variations. For face images under pose 45° and -45° , over 35% improvement is achieved on the recognition rates.

B. Labeled Faces in the Wild

LFW [34] contains 13 223 images from 5749 subjects with various poses, expressions, illuminations, etc. Our architecture is trained on CelebFaces [37], in which the images are with large pose variations.

1) *Selection of Frontal Face Images*: We adopt the same strategy as [36], which also recovers frontal faces for face verification task. The measurement of the frontal view face image is defined by

$$M(X_i) = \|X_i L - X_i R\|_F^2 - \lambda \|X_i\|_* \quad (5)$$

where X_i denotes a face image of the i th identity, λ is a constant, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_*$ denotes the nuclear norm. L and R are two constant matrices representing the left part and right part of the input face image, respectively. The first term in (5) corresponds to the symmetry of the face, while the second represents face rank.

TABLE IV
FACE VERIFICATION PERFORMANCE (%) USING DIFFERENT FEATURES ON THE LFW DATASET

Methods	LBP	Gabor	HOG
Zhu [36]	87.31	85.17	83.26
TSDCN	87.59	87.34	88.48

TABLE V
CLASSIFICATION ACCURACIES (%) OF DIFFERENT POSES ON THE MULTI-PIE DATASET

View	-45°	-30°	-15°	$+15^\circ$	$+30^\circ$	$+45^\circ$	Avg
Accuracy	96.1	95.2	95.3	96.7	89.4	93.2	94.3



Fig. 6. Reconstructed examples on the LFW dataset: (left) Original face images. (right) Reconstructed frontal face images under neutral illumination.

TABLE VI
COMPARISONS OF RECOGNITION RATES (%) UNDER DIFFERENT POSES USING THE CONCATENATION LAYER AND TSDCN ON THE MULTI-PIE DATASET

	-45°	-30°	-15°	$+15^\circ$	$+30^\circ$	$+45^\circ$	Avg
Concat	57.2	81.5	90.0	94.0	80.7	52.5	76.0
TSDCN	69.2	88.9	93.3	95.5	89.0	69.6	84.3

2) *Evaluation on the Recovered Face Images*: To evaluate the quality of the reconstructed face images, HOG [38], LBP [39], and Gabor [40] features are extracted from the recovered frontal images on the LFW dataset. PCA is applied after feature extraction. Table IV reports the performance of face verification on the LFW dataset, it demonstrates that our method performs favorably against Zhu [36]. For HOG feature, the accuracy is improved by 5.22% using our method. Fig. 6 shows some recovered face images with identity preserving.

C. Effectiveness of the Correlation Layer

If the correlation layer is replaced by the concatenation layer in the TSDCN, the correlation between the geometric information and textural information is completely abolished. Table VI indicates that our correlation layer achieves higher recognition rates than concatenation layer. Best results are written in bold. Meanwhile, we observe that the improvement over concatenation layer integration is more obvious when the input face is in larger pose.

IV. CONCLUSION

In this letter, we address the frontal face recovery problem through a TSDCN. Different from existing methods only considering texture correspondence, the proposed network incorporates both geometric and textural features to construct a unified network. The overall performance of TSDCN is superior compared with state-of-the-art frontal face recovery methods. Owing to the specific capability to perform accurate face geometry modeling, we obtain reconstructed frontal face images of high quality. In the future work, we will further extend our TSDCN architecture for generic image rotating task.

REFERENCES

- [1] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 113–120.
- [2] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.
- [3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1891–1898.
- [4] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
- [5] O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. Brit. Mach. Vision Conf.*, vol. 41, 2015, pp. 1–12.
- [6] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 365–372.
- [7] Y. Taigman *et al.*, "Multiple one-shots for utilizing class label information," in *Proc. Brit. Mach. Vision Conf.*, vol. 2, 2009, pp. 1–12.
- [8] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 498–505.
- [9] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 497–504.
- [10] C. Huang, S. Zhu, and K. Yu, "Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval," NEC, Princeton, NJ, USA, Tech. Rep. TR115, 2011.
- [11] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 676–684.
- [12] S. R. Arashloo and J. Kittler, "Energy normalization for pose-invariant face recognition based on mrf model image matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1274–1280, Jun. 2011.
- [13] T. Zhang, Q. Dong, and Z. Hu, "Pursuing face identity from view-specific representation to view-invariant representation," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3244–3248.
- [14] Y. Li and J. Feng, "Frontal face synthesizing according to multiple non-frontal inputs and its application in face recognition," *Neurocomputing*, vol. 91, pp. 77–85, 2012.
- [15] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3499–3506.
- [16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2539–2547.
- [17] J. Yang, S. E. Reed, M. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1099–1107.
- [18] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognit.*, vol. 42, no. 11, pp. 2876–2896, 2009.
- [19] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 937–944.
- [20] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 102–115.
- [21] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4295–4304.
- [22] L. Chang and D. Y. Tsao, "The code for facial identity in the primate brain," *Cell*, vol. 169, no. 6, pp. 1013–1028, 2017.
- [23] A. Dosovitskiy *et al.*, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
- [25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [27] Y. Cheng, X. Zhao, R. Cai, Z. Li, K. Huang, and Y. Rui, "Semi-supervised multimodal deep learning for rgb-d object recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3345–3351.
- [28] A. Li, S. Shan, and W. Gao, "Coupled bias-variance tradeoff for cross-pose face recognition," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 305–315, Jan. 2012.
- [29] Z. Wu and W. Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2016, pp. 1–6.
- [30] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recognit.*, vol. 66, pp. 144–152, 2017.
- [31] X. Zhao *et al.*, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 425–442.
- [32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. IEEE ACM Int. Conf. Multi.*, 2014, pp. 675–678.
- [33] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
- [34] 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," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
- [35] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [36] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Recover canonical-view faces in the wild with deep neural networks," 2014, arXiv preprint arXiv:1404.3543.
- [37] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.
- [38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
- [39] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
- [40] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, no. 7, pp. 1160–1169, 1985.