RESIDUAL LEARNING FOR FACE SKETCH SYNTHESIS

Junjun Jiang[1,2]*, Yi Yu*[1]*, Zheng Wang*[1]*, and Jiayi Ma*[3]

[1] National Institute of Informatics, Tokyo, Japan
[2] China University of Geosciences, Wuhan, China
[3] Wuhan University, Wuhan, China

ABSTRACT

Face sketch synthesis plays an important role in both digital entertainment and law enforcement. It can bridge the great texture discrepancy between face photos and sketches. Most of the current face sketch synthesis approaches directly learn the relationship between the photos and sketches, and it is very difficult for them to generate the individual specific details, which we call *rare features*. To address this problem, in this paper we propose a novel face sketch synthesis through residual learning. In contrast the traditional approaches, which try to construct the sketch image directly, we aim at predicting the residual image (between the photo and sketch), given the photo observation. In addition, we also introduce a couple dictionary learning algorithm through preserving the local geometry structure of data space, which is usually ignored by existing methods. Our proposed method shows impressive results on the face sketch synthesis task, when compared with some state-of-the-arts including some recent proposed deep learning based approaches.

Index Terms— Face sketch synthesis, residual learning, dictionary learning, locality-constrained representation

1. INTRODUCTION

Face sketch synthesis refers to infer a face sketch from a facial photo given some photo and sketch training pairs, which can be seemed as one of the cross-modal image transformation problem [1, 2]. It is a very active research topic because it can narrow the gap between different modalities, which is the biggest obstacle to cross-modal face recognition [3, 4]. For example, surveillance cameras have been widely used in security and protection systems. They can provide very important clues about objects for solving a case, such as criminals [5, 6]. However, the object is so far away from the camera that the resolution of the interested face in the picture is too low to provide helpful information [7, 8]. More generally, the surveillance camera system may not shoot the suspect's face image at all, and there is no information other than the witnesses' descriptions. A sketch drawn by the artist is usually taken as the substitute for suspect identification [9]. Therefore, face sketch synthesis can be seen as a tool that bridges the great texture discrepancy between face sketches and photos [10, 11, 12].

(a) Sketch Image (b) Photo Image (c) Residual Image

Fig. 1. Illustration of (c) residual image by subtracting (b) photo image from (a) sketch image. Given a photo image, traditional face sketch synthesis approaches try to directly predict the sketch image, while our proposed method aims at predicting the less specific residual image.

Prior Work. Thanks to the rapid development of machine learning technologies, there are many learning-based face sketch synthesis methods have been developed since the pioneer work of [13, 14, 1]. They can overcome the problem that traditional image-based methods cannot well mimic the sketch style and their reconstructed sketches are more like photos [15, 16]. Similar to these manifold learning based face hallucination methods [17, 18], the common idea of these learning-based approaches is to embed the local manifold structure of photo/image patch space to that of the image/photo space based on the manifold assumption which states that the coupled spaces share the similar local geometry structure [19, 20]. Specially, given an input photo image (here we take face sketch synthesis as an example, the same as the photo sketch synthesis), they firstly divide the photo to small patches. For each testing photo patch, they search K most similar patches from the photo patch training space, and they obtain the optimal combination coefficients by minimizing the linear reconstruction error of the testing photo patch with its K nearest neighbors or sparse neighbors. Based on

The research was supported by the National Natural Science Foundation of China under Grant 61501413 and Grant 61502354, and the JSPS KAKENHI Grant Number 16K16058..

the manifold assumption, the output of sketch image can be predicted with the corresponding K most similar patches or sparse neighbors from the sketch patch training space and the same optimal combination coefficients. In order to comply the compatibility between neighbor patches, Markov random fields (MRF) [1, 21] and Markov weight fields (MWF) [22] based regularizations have been widely used. Most recently, some methods that benefit from the deep learning technologies have emerged, such as GAN [23], deep graphical feature learning [24], and fully convolution networks [25, 26, 27].

Motivations. However, when the number of training samples is limited, it is very difficult for these learning-based methods to predict the individual specific details of the input image, which we call *rare features* in this paper. Therefore, in this paper we proposed a novel face sketch synthesis framework through residual learning. As some of the rare features is eliminated by subtraction (as shown in Fig. 1), this greatly reduces the difficulty of learning and prediction. Meanwhile, in order to predict the residual images more accurately, we further develop a local geometry constrained photo-sketch (residual) couple dictionary learning algorithm, thus transforming the face sketch synthesis from the image space to a new and compact space spanned by the learned dictionary atoms, where the manifold assumption can be better guaranteed. Experimental results on three public photo-sketch face databases show the superiority of the proposed method over some state-of-the-art including some recent proposed deep learning based approaches.

2. THE PROPOSED METHOD

Observing that the sketch differences of individuals are narrowed by subtracting the photo image from the corresponding sketch image, in the paper we propose a novel face sketch synthesis framework based on residual learning, which will make it much easier to predict the rare sketch features for an input photo. In the following, we will present the details of the proposed method. Note that when we mention the sketch dictionary or sketch images, we mean their residual ones to make the statement more concise and convenient.

We first introduce some notations used in this paper. The aim of face sketch synthesis is to predict the sketch $\mathbf{x}_t^s$ of an observed photo $\mathbf{x}_t^p$, given N pairs of training photos and training sketches, $\mathbf{X}^p$ and $\mathbf{X}^s$. Here, the subscript "t" is used to indicate the test image, the superscripts "p" and "s" are used to indicate the photo and sketch respectively. As for a local patch based approach, we divide the input photo, the photo and sketch images into small patches according to the position, $\mathbf{x}_{t(i,j)}^p$, $\mathbf{X}_{(i,j)}^p$ and $\mathbf{X}_{(i,j)}^s$, respectively, where the term (i, j) indicates the location of the patch at the i-th row and the j-th column. As discussed above, we extend the training samples by incorporating all the candidate patches around the position patch. In order to make the expression more concise, we directly use $\mathbf{X}_{(i,j)}^p$ and $\mathbf{X}_{(i,j)}^s$ to denote the extended training patches. If it does not lead to misunderstanding, we also drop the term (i, j) in the following.

The main idea of patch based methods is to transform the patch representations from the photo (observation) space to the sketch (target) space. Thus, the key problem is how to obtain the optimal representations of an input photo patch. K nearest neighbor and sparse neighbor are the most widely used constraints for patch representation [19, 20]. However, the former will induce the over- or under-fitting problem, while the latter cannot exploit the locality prior of the data space.

To preserve the local geometry structure of data space, we introduce a local manifold geometry constrained representation method. In particular, the representation $\mathbf{a}_t^p$ of an input photo patch $\mathbf{x}_t^p$ can be obtained by minimizing the following reconstruction errors:

$$\begin{aligned} \mathbf{a}_t^p &= \arg\min_{\mathbf{a}_t^p} \|\mathbf{x}_t^p - \mathbf{D}^p\mathbf{a}_t^p\|_F^2 + \lambda\|\mathbf{c}_t \odot \mathbf{a}_t^p\|_2^2 \\ \text{s.t.} \quad & \mathbf{1}^T\mathbf{a}_t^p = 1, \end{aligned} \tag{1}$$

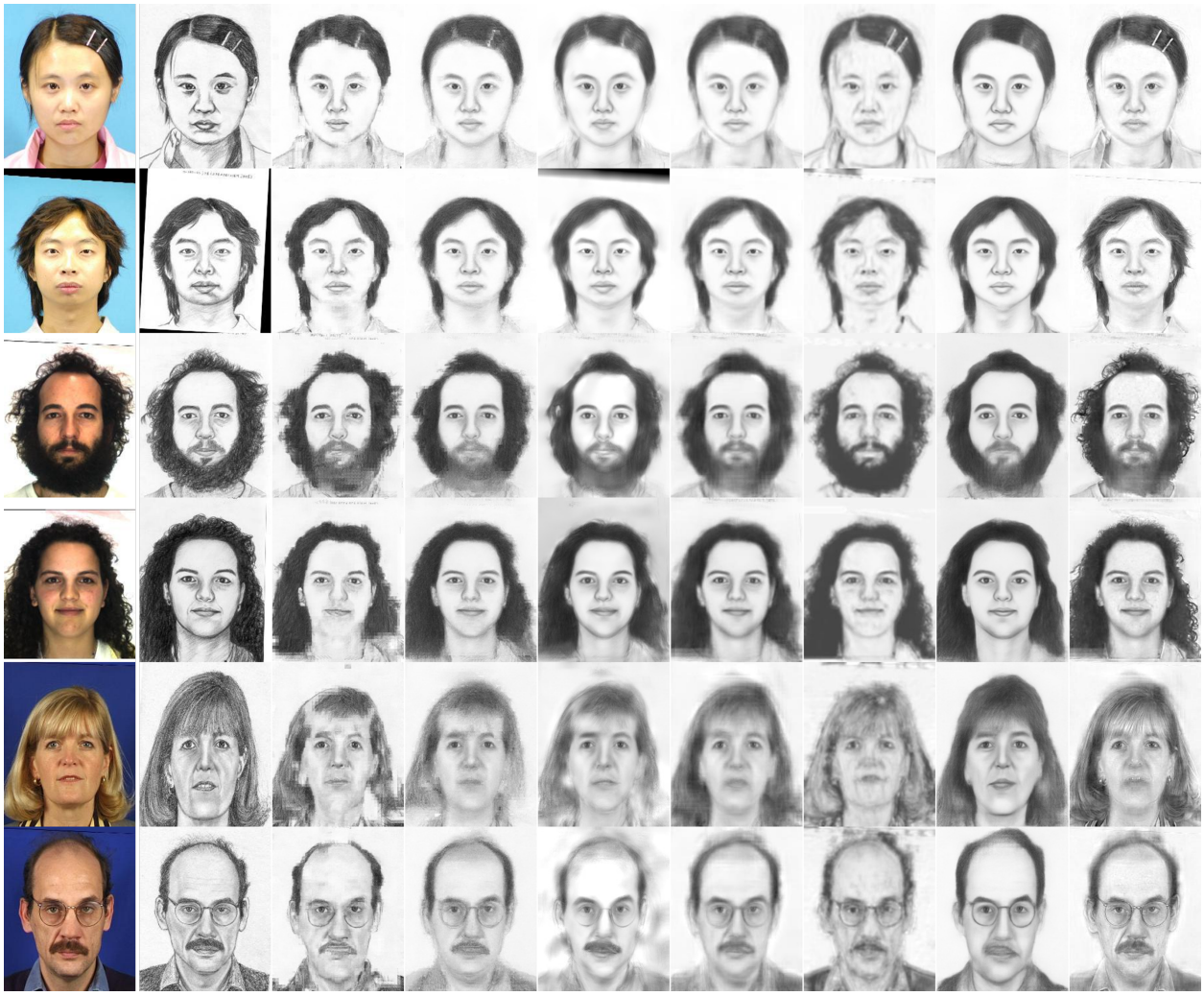
where $\odot$ denotes the element-wise multiplication, $\mathbf{c}_t$ is a $K \times 1$ vector and denotes the locality adaptor that gives different freedom for each basis element proportional to its similarity to the input sample $\mathbf{x}_t^p$, $c_{kt} = ||\mathbf{x}_t^p - \mathbf{d}_k^p||_2, k = 1, 2, ..., K$, the parameter λ is used to balance the contribution between the reconstruction error and locality prior. The sum-to-one constraint follows the shift-invariant requirements of [28]. Here, we assume that we have learned the photo-sketch couple dictionary, $\mathbf{D}^p$ and $\mathbf{D}^s$, with K elements.

Acquiring the optimal photo patch representation $\mathbf{a}_t^p$, the target sketch patch $\mathbf{x}_t^s$ can be generated by the linear combination of the sketch dictionary and the representation obtained by the observed photo patch:

$$\mathbf{x}_t^s = \mathbf{D}^s\mathbf{a}_t^p \tag{2}$$

2.1. Locality-constrained Photo-Sketch Couple Dictionary Learning

In all the above discussions, we have assumed that the dictionary is given. Recently, sparse dictionary learning has been widely used to recover the underlying structure in many kinds of natural data. Instead of working directly with the image patches sampled from the training database, it learns a compact dictionary and representation for these patches to capture the sparsity prior, significantly improving the speed of the algorithm and reconstruction and analysis accuracy [31, 32]. However, algorithms of this type do not preserve the information of data locality, i.e., two very similar image patches may get two completely different representations. This is not what we expect because it cannot faithfully depict intrinsic manifold geometry of data space. It has been pointed out in [28] that locality is more essential than sparsity, as locality would imply sparsity but not necessarily vice versa. By incorporating the locality constraint, it can achieve locality and sparsity



(a) Photo (b) Sketch (c) MRF [1] (d) MWF [22] (e) SSD [29] (f) RSLCR [30] (g) FCN [25] (h) GAN [23] (i) Our

Fig. 2. Four groups of face sketch synthesis results on three databases.

simultaneously [33]. This motivates us to propose a locality-constrained photo-sketch couple dictionary learning algorithm which can guarantee performance and convergence.

Similar to [34], we decompose the couple dictionary learning problem to the dictionary optimization of one domain (i.e., the photo training samples) and performing pseudo inverse for the other domain (i.e., the sketch training samples) while preserving the same representations. The problem of learning a photo patch dictionary $\mathbf{D}^p$ for locality representation, in its most popular form, is solved by minimizing the energy function that combines squared reconstruction errors and the locality penalties $\mathbf{C}$ on the representations $\mathbf{A} = [\mathbf{a}_1^p, \mathbf{a}_2^p, ..., \mathbf{a}_N^p]$:

$$(\mathbf{D}^p, \mathbf{A}) = \arg\min_{\mathbf{D}^p, \mathbf{A}} \|\mathbf{X}^p - \mathbf{D}^p\mathbf{A}\|_F^2 + \lambda \sum_{i=1}^{N} \|\mathbf{c}_i \odot \mathbf{a}_i\|_2^2$$
$$\text{s.t.} \quad \mathbf{1}^T\mathbf{a}_i = 1, \quad \forall\, i = 1, 2, \cdots, N. \tag{3}$$

where $\mathbf{c}_i$ is a $K \times 1$ vector that gives different freedom for each basis element proportional to its similarity to the input sample $\mathbf{x}_i^p$. The objective function of (3) can be solved iteratively optimizing $\mathbf{D}^p(\mathbf{A})$ based on existing $\mathbf{A}(\mathbf{D}^p)$ [28].

The next step is the sketch patch dictionary $\mathbf{D}^s$ construction, given the photo patch dictionary $\mathbf{D}^p$ and the representations $\mathbf{A}$. Based on the assumption that the photo-sketch pairs share the same representations $\mathbf{A}$, the sketch patch dictionary $\mathbf{D}^s$ is defined to be the one that minimizes the mean approximation error, i.e.,

$$\mathbf{D}^s = \arg\min_{\mathbf{D}^s} \|\mathbf{X}^s - \mathbf{D}^s\mathbf{A}\|_F^2 . \tag{4}$$

The solution of the problem (4) is given by the following Pseudo-Inverse expression (given that $\mathbf{A}$ has full row rank):

$$\mathbf{D}^s = \mathbf{X}^s\mathbf{A}^+ = \mathbf{X}^s\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}. \tag{5}$$

3. EXPERIMENTAL RESULTS

In this section, we describe the details of experiments performed to evaluate the effectiveness of the proposed shallow residual learning based framework for face sketch synthesis. We compare our method with several recent state-of-the-art methods including the MRF [1], the MWF [22], spatial sketch denoising (SSD) [29], random sampling and locality constraint for face sketch synthesis (RSLCR) [30], and two deep learning based methods, the fully convolutional network (FCN) based method [25] and GAN based method [23]. The experiments are carried out on the public CUHK database [1] (88 samples for training and the remaining 100 for testing), AR database [35] (80 samples for training and the remaining 43 for testing), and XM2VTS database [36] (100 samples for training and the remaining 195 for testing). In our experiments, all face images are cropped into 250×200. Similar to previous work, we also divide the databases into training samples and testing samples.

Table 1. The face recognition results of comparison methods with different reduced dimensions by NLDA. The best and second best results are marked in red and blue, respectively.

Dim	5	10	20	50	100	149
MRF [1]	44.97	67.15	77.95	84.23	86.81	87.34
MWF [22]	52.07	72.47	82.39	89.34	91.81	92.13
SSD [29]	47.37	68.70	78.22	86.84	90.21	90.35
RSLCR [30]	70.64	87.10	93.27	96.70	97.71	98.06
FCN [25]	66.54	85.61	91.54	95.05	96.22	96.62
GAN [23]	64.04	82.26	89.15	92.23	92.93	93.16
Our	73.30	88.61	93.89	97.09	97.82	98.18

Fig. 2 shows the results of face sketch synthesis with different methods. The first and second columns are the input photos and original sketches. From the third column to the last column, they are the results of four shallow learning based methods (MRF [1], MWF [22], SSD [29], and RSLCR [30]), two deep learning based methods (FCN [25]and GAN [23]), and our proposed method. These shallow learning based methods either fail to obtain clear contours (please refer to the results of MRF [1] and MWF [22]) or are too smooth (please refer to the results of SSD [29] and RSLCR [30]). FCN [25] seems can synthesize some detailed features, but will also introduce some unexpected distortions. The results of GAN [23] are sharp and look very pleasant, but they lost the rare details, such as the headdress of the first person and the hair of the second and third person. By predicting the residuals (instead of the sketch), our approach avoids the difficulty of directly predicting the individual specific and rare features of sketch images. As shown in the results, our method can well predict the rare features while the comparison methods tend to smooth these regions. Please carefully check the headdress, hair, glasses and other details.

Following [20, 30, 23], we randomly select 150 synthesized sketches and the corresponding ground-truth sketches to train the classifier. The remaining 188 sketches are used as the gallery. We repeat the experiment 20 times by randomly dividing. Table 1 shows the recognition rates versus feature with different dimensions by NLDA [37]. It should be noted that as an LDA based method, the maximum reduced dimensionality of NLDA is equal to $C-1$, where C is the class number and is set to 150 in our experiments. Our method could achieve the best performance under various dimensions. GAN based method [23] can get a good visual result, but its face recognition rate is relatively poor. This is mainly because that sharp edge of reconstructed sketch images by GAN [23] do not mean high fidelity.

4. CONCLUSIONS

In this paper, we advocated a novel framework toward face sketch synthesis with residual learning. We observed that it is much easier for a learning based method to predict the individual specific and rare features of the target sketch image from the residual sketches than from the original sketch images. In addition, we also proposed a novel photo-sketch couple dictionary learning approach through exploiting the local manifold geometry of data space. Experiments demonstrate the superiority of our method when compared with some state-of-the-art including some recent proposed FCN [25] and GAN [23] based deep learning approaches, especially when the input photo has some specific features that rarely appear in the training database.

5. REFERENCES

[1] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[2] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *International Journal of Computer Vision*, vol. 106, no. 1, pp. 9–30, 2014.

[3] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, Afzel Noore, and Angshul Majumdar, "Face sketch matching via coupled deep transform learning," in *ICCV*, 2017.

[4] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li, "A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution," *arXiv preprint arXiv:1409.5114*, 2014.

[5] Z. Wang, R. Hu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *IJCAI*, 2016, pp. 2669–2675.

[6] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans Cybern*, vol. PP, no. 99, pp. 1–15, 2017.

[7] J. Jiang, Y. Yu, S. Tang, J. Ma, G. Qi, and A. Aizawa, "Context-patch based face hallucination via thresholding locality-constrained representation and reproducing learning," in *ICME*, 2017, pp. 469–474.

[8] J. Jiang, R. Hu, Z. Han, T. Lu, and K. Huang, "A super-resolution method for low-quality face image through RBF-PLS regression and neighbor embedding," in *ICASSP*, 2012, pp. 1253–1256.

[9] Fei Gao, Shengjie Shi, Jun Yu, and Qingming Huang, "Composition-aided sketch-realistic portrait generation," *arXiv preprint arXiv:1712.00899*, 2017.

[10] Shuxin Ouyang, Timothy M. Hospedales, Yi-Zhe Song, and Xueming Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *CVPR*, 2016, pp. 5571–5579.

[11] Yibing Song, Jiawei Zhang, Linchao Bao, and Qingxiong Yang, "Fast preprocessing for robust face sketch synthesis," in *IJCAI*, 2017, pp. 3574–3580.

[12] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1586–1590, Nov 2017.

[13] X. Tang and X. Wang, "Face photo recognition using sketch," in *ICIP*, 2002, vol. 1, pp. I–257–I–260 vol.1.

[14] Xiaoou Tang and Xiaogang Wang, "Face sketch recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50–57, 2004.

[15] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *CVPR*. IEEE, 2012, pp. 2216–2223.

[16] D. A. Huang and Y. C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *ICCV*, Dec 2013, pp. 2496–2503.

[17] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. on Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug 2014.

[18] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct 2014.

[19] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma, "A nonlinear approach for face sketch synthesis and recognition," in *CVPR*. IEEE, 2005, vol. 1, pp. 1005–1010.

[20] Xinbo Gao, Nannan Wang, Dacheng Tao, and Xuelong Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 8, pp. 1213–1226, 2012.

[21] Wei Zhang, Xiaogang Wang, and Xiaoou Tang, "Lighting and pose robust face sketch synthesis," in *ECCV*. Springer, 2010, pp. 420–433.

[22] Hao Zhou, Zhanghui Kuang, and Kwan-Yee K Wong, "Markov weight fields for face sketch synthesis," in *CVPR*. IEEE, 2012, pp. 1091–1097.

[23] Nannan Wang, Wenjin Zha, Jie Li, and Xinbo Gao, "Back projection: An effective postprocessing method for gan-based face sketch synthesis," *Pattern Recognition Letters*, 2017.

[24] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li, "Deep graphical feature learning for face sketch synthesis," in *IJCAI (IJCAI-17)*, 2017, pp. 3574–3580.

[25] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *ICMR*. ACM, 2015, pp. 627–634.

[26] Dongyu Zhang, Lin Liang, Tianshui Chen, Wu Xian, Wenwei Tan, and Ebroul Izquierdo, "Content-adaptive sketch portrait generation by decompositional representation learning," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 26, no. 1, pp. 328–339, 2016.

[27] Chaofeng Chen, X Tax, and KKY Wong, "Face sketch synthesis with style transfer using pyramid column feature," in *IEEE Winter Conference on Applications of Computer Vision*. Lake Tahoe, USA, 2018.

[28] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *NIPS*, 2009, pp. 2223–2231.

[29] Yibing Song, Linchao Bao, Qingxiong Yang, and Ming-Hsuan Yang, "Real-time exemplar-based face sketch synthesis," in *ECCV*. Springer, 2014, pp. 800–813.

[30] Nannan Wang, Xinbo Gao, and Jie Li, "Random sampling and locality constraint for face sketch," *CoRR*, vol. abs/1701.01911, 2017.

[31] J. Yang, J. Wright, Thomas H., , and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[32] J. Jiang, R. Hu, Z. Han, Z. Wang, and J. Chen, "Two-step superresolution approach for surveillance face image through radial basis function-partial least squares regression and locality-induced sparse representation," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 041120, 2013.

[33] J. Jiang, R. Hu, Z. Han, T. Lu, and K. Huang, "Position-patch based face hallucination via locality-constrained representation," in *ICME*, 2012, pp. 212–217.

[34] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[35] A. M. Martinez and R. Benavente, "The AR face database," *Computer Vision Center, Barcelona, Spain, Tech. Rep. 24*, Jun. 1998.

[36] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, 1999, vol. 964, pp. 965–966.

[37] Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.