LANDMARKGAN: SYNTHESIZING FACES FROM LANDMARKS

Pu Sun[1*], *Yuezun Li*[2*], *Honggang Qi*[1] *and Siwei Lyu*[2]

[1] University of Chinese Academy of Sciences, China
[2] University at Buffalo, State University of New York, USA

ABSTRACT

Face synthesis is an important problem in computer vision with many applications. In this work, we describe a new method, namely LandmarkGAN, to synthesize faces based on facial landmarks as input. Facial landmarks are a natural, intuitive, and effective representation for facial expressions and orientations, which are independent from the target's texture or color and background scene. Our method is able to transform a set of facial landmarks into new faces of different subjects, while retains the same facial expression and orientation. Experimental results on face synthesis and reenactments demonstrate the effectiveness of our method.

***Index Terms*—** Face synthesis, GAN

1. INTRODUCTION

Creating realistic images of human faces, as an important problem in computer vision with many practical applications, has recently received a lot of attentions [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. The general approach of face synthesis is to use a generator, usually in the form of a deep neural network, which takes an input control variable and converts it to a face image. The early face synthesis methods [1, 2, 10] are based on Generative Adversarial Networks (GANs) [15], which use random noise as the input control methods. Although highly realistic human face images are generated using these methods, they have a major limitation: the user has little control over the identity and facial attributes such as expression and orientation in the synthesized faces. Face style transfer methods [3, 4, 5, 6, 7, 11] generate new face images by incorporating the style transferred from other domain to the source face image instead of input random noise. Subsequently, face reenactment methods [8, 9, 12, 13] take face images of a source identity as the system input, and generate faces of a different target identity preserving the facial expression of the input.

In this work, we describe a new method, known as LandmarkGAN, to synthesize faces only using facial landmarks. Facial landmarks correspond to important locations of facial parts (tips and middle points of eyes, nose, mouth, and eye brows) and contours. The facial landmarks can be reliably detected from input images using state-of-the-art algorithms [16, 17]. As the extraction of facial landmarks discard texture and color of the individual faces and any non-face backgrounds, they provide a natural low-dimensional representation for synthesizing faces. Moreover, the facial landmarks are structural and human interpretable compared to other signals such as Action Units (AU) [18, 19], which enables direct editing to generate faces with modified expressions and orientations. Note many previous works achieve the face synthesis or reenactment assisted with the guidance such as facial landmarks, yet few of them focus on synthesizing faces solely based on facial landmarks. The work [20] synthesizes faces from facial landmarks. However, it only focuses on persevering the genders of generated images instead of identity switching, facial expressions and orientations, which therefore hardly to be applied in reenactment. Fig.1 shows two groups of visual examples generated by our method. We edit the facial landmarks to synthesize corresponding target faces[1]. The target identity is randomly selected from CelebV dataset [9].

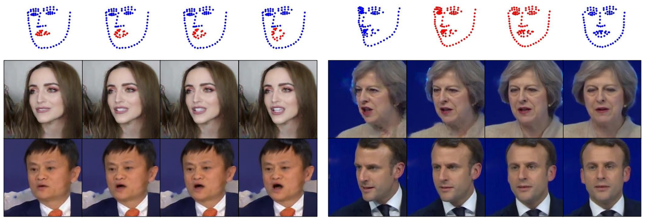

Fig. 1. Visual examples generated by our method. The first row is the input facial landmarks with mouth and head orientation edited (red marks in left and right figure) and the other two rows are synthesized faces of target identity using facial landmarks as input, which showcase the facial expressions and orientations are greatly retained.

The proposed face synthesis model has two components. The first is a *landmark converter*, which takes an auto-encoder structure to convert the input facial landmarks to those of the target. The converted facial landmarks are then fed to a *target-specific landmark-to-face (TL2F) generator*, which is an up-sampling convolutional neural network, to create the face image of target identity incorporating the facial expression and orientation in the facial landmarks. We then describe a new fully differentiable landmark detector to en-

[*] The authors contribute equally.

[1] We provide a GUI for editing faces. The demo can be found here `https://drive.google.com/file/d/1gd_vgCqEULeWt4DMjvXuqqtLTxNQMY6n/view?usp=sharing`.

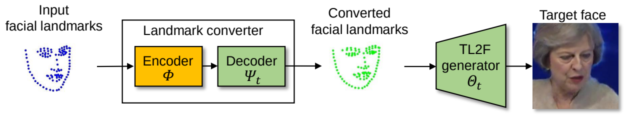

Fig. 2. Overview of our method to synthesize a target face.

able landmark transferring consistently. Our model is trained by jointly optimizing the parameters in landmark converter and TL2F generator.

Experimental results show that our method can synthesize face images of target identity with high visual quality and varying facial expressions and orientations. We also implement a face reenactment system based on our method, where the input facial landmarks are extracted from the input faces. When compared with state-of-the-art face reenactment methods, our method achieves competitive performance with improved qualitative and quantitative evaluation results.

2. METHODOLOGY

2.1. Model Structure

The overall structure of our method is illustrated in Fig.2. When the input is the face of a source identity, one can use any off-the-shelf facial landmark detection methods, such as [21, 16] to extract landmarks and use them as inputs for landmark-based face reenactment. In the following, we use Ł_s and $\mathbf{I}_s$ to denote the input facial landmarks and corresponding image of source identity s, respectively.

The (x,y)-coordinates of the input facial landmarks are first converted to a vector and input to the *landmark converter*. The landmark converter has an auto-encoder structure, which is formed by an encoder and a decoder, both are lightweight neural networks consisting of five fully connected (FC) layers. The encoder, subsequently denoted as Φ, is shared across all identities, and converts the input landmark into a latent feature that is identity-neutral but preserve essential facial expressions and orientations. On the other hand, the decoder is *target-specific*, which reconstructs landmarks specific to the shape and geometry of its corresponding target from the latent feature. We subsequently denote the decoder in the landmark converter as Ψ_t for target t. With a set of facial landmarks for source s, Ł_s, as input, the converted facial landmarks of target t, $t \neq s$ is obtained as $\bar{\text{Ł}}_t = \Psi_t(\Phi(\text{Ł}_s))$.

The converted landmarks are then used as the input to a *target-specific landmark-to-face (TL2F) generator*, which synthesizes the face image of target identity corresponding to the facial expression and head orientations represented in the input facial landmarks. The TL2F generator consists of two FC layers and six upscale blocks. The FC layers transform the converted landmarks to a feature vector, which is then reshaped to a feature map. The upscale block is a set of operations that upsamples the input feature map by scale 2 in width and height. In detail, the upscale block contains a convolutional layer which increases the channel of input feature map, and a PixelShuffle layer [22] which upsamples the feature map by shifting the elements in channel dimension to width and height dimension. In what follows, we will use Θ_t to denote the TL2F generator for target t. With the converted facial landmarks $\bar{\text{Ł}}_t$, the face image of target t synthesized by Θ_t is represented as $\bar{\mathbf{I}}_t = \Theta_t(\bar{\text{Ł}}_t)$.

2.2. Training

As we do not assume correspondence in facial expressions among landmarks of different identity, the loss function is formed in a self-regularized manner. The overall loss function is the sum of five loss terms, as

$$L_{\text{overall}} = L_{\text{L2I}} + L_{\text{I2L}} + L_{\text{L2L}} + L_{\text{X-L2L}} + L_{\text{L2I-gan}}. \tag{1}$$

The first term corresponds to the ℓ_1 error between an input image and its reconstruction from its landmarks using the target specific landmark-to-face generator. Specifically, for a target face t with an input image $\mathbf{I}_t$ and the corresponding facial landmarks Ł_t, the image reconstruction loss is given by

$$L_{\text{L2I}} = \mathbb{E}_t[\|\mathbf{I}_t - \Theta_t(\text{Ł}_t)\|_1], \tag{2}$$

where $\mathbb{E}_t$ denotes the average over all training identities.

The second term ensures the facial landmarks are well preserved in synthesized face. Denote Ω as the landmark detector. This term can be written as

$$L_{\text{I2L}} = \mathbb{E}_t[\ ||\text{Ł}_t - \Omega(\Theta_t(\text{Ł}_t))||_2\ ]. \tag{3}$$

Differentiable landmark detector. To be able to optimize this loss, a differentiable landmark detector is required. However, this is not the case for most existing state-of-the-art landmark detectors [21, 16, 17], due to the non-differentiable argmax function used for final landmark selection. Therefore, we adapt the *differentiable spatial to numerical transform* (DSNT) module [23] to make a differentiable landmark detector. The DSNT converts the non-differentiable argmax function into a soft-argmax [24] based function which can be differentiable.

The third term corresponds to the reconstruction of landmarks in the landmark converter, as

$$L_{\text{L2L}} = \mathbb{E}_t[\ ||\text{Ł}_t - \Psi_t(\Phi(\text{Ł}_t))||_2 + ||\text{Ł}_t^{'} - \Phi(\text{Ł}_t)||_2\ ], \tag{4}$$

where the first term of Eq.(4) measures the landmark error between the input facial landmarks Ł_t and its reconstruction from the landmark converter, combining the general encoder Φ and target t's decoder Ψ_t, $\text{Ł}_t'$ in second term is the ℓ_2 normalized landmarks Ł_t, which guides the encoder to learn identity-independent landmarks in auxiliary.

The four term evaluates errors when landmarks of different identities' faces undergo with the landmark converter. Specifically, it is the difference between the landmark of identity s and its cyclic transformation to and back from identity t, as

$$L_{\text{X-L2L}} = \mathbb{E}_{t \neq s}[\ ||\text{Ł}_t - \Psi_t(\Phi(\Psi_s(\Phi(\text{Ł}_t))))||_2\ ]. \tag{5}$$

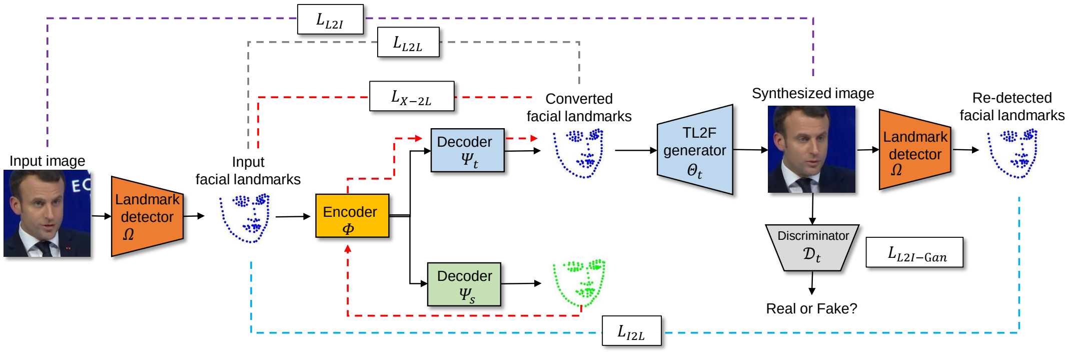
L_{L2I}
L_{L2L}
L_{X-2L}
Input image
Landmark detector Ω
Input facial landmarks
Encoder Φ
Decoder Ψ_t
Decoder Ψ_s
Converted facial landmarks
TL2F generator Θ_t
Synthesized image
Landmark detector Ω
Re-detected facial landmarks
Discriminator $\mathcal{D}_t$
$L_{L2I-Gan}$
Real or Fake?
L_{I2L}

Fig. 3. Overview of training process of our method. See text for details.

The last term couples the training of the landmark converter and the target specific landmark to image generator, which is a landmark-to-image GAN loss as

$$L_{\text{L2I-gan}} = \begin{array}{l} \mathbb{E}_t[[\,\log \mathcal{D}_t(\mathbf{I}_t)\,]+ \\ \mathbb{E}_s[\,\log(1-\mathcal{D}_t(\Theta_t(\Psi_t(\Phi(\text{Ł}_s)))\,]], \end{array} \tag{6}$$

where $\mathcal{D}_t$ denotes a discriminator as in the PatchGAN model [25] to distinguish the original and synthesized face images of identity t.

3. EXPERIMENTS

3.1. Experimental Settings

Dataset. We train and test our method using the CelebV dataset [9]. We choose this dataset because it has been used in previous works [9, 13]. The CelebV dataset includes faces of five identities, namely, Emmanuel Macron, Kathleen, Jack Ma, Theresa May, and Donald Trump.

Implementation details. Our method is implemented using PyTorch 1.0.1 on Ubuntu 16.04 with a Nvidia 1080ti GPU. The models are trained using the RMSProp optimizer [26] with kaiming initialization [27] of the model parameters. For the landmark converter, the training batch size is 4, the learning rate starts as 10^{-5}, and the maximum iteration is set to $45,000$. For the target-specific landmark-to-image generator, the batch size is set to 1 and the maximum iteration is set to 4×10^5. The learning rate starts as 6×10^{-5}, and is decayed 10% every $2,500$ iterations. For the differentiable landmark detector, we use one stacked hourglass structure as the base network to save the resource cost in training. Our landmark detector is trained on the WFLW dataset [28].

3.2. Landmark to Face Synthesis

Fig. 1 shows several examples of landmark to face synthesis. To further demonstrate the flexibility of our method, we conduct another set of experiments to progressively change the landmarks corresponding to eyes, see Fig.4 (a). We further generate faces with more extreme editing of landmarks to see the response of our method. The first setting is we shift the location of mouth to an extreme location that is not existed in the training data in CelebV. In the second setting, we edit the face contour to change the shapes of the targets' faces. Fig.4 (b) shows the visual examples of our method to different inputs, where the top and bottom part correspond to two settings respectively. As in the previous cases, we can observe that the edited facial landmarks lead to corresponding changes in the synthesized faces. Moreover, our method can adjust the synthesized face properly to keep the semantic meaning even though the shapes of the faces are changed. These confirm that our method are flexible to larger changes to the landmarks and can synthesize faces with variations that are not present in the training data. However, our method still has the limit which can not handle the facial landmarks with extreme editing such as moving the mouth outside of face or largely shrinking the face contour. Fig.4 (c) shows several failure examples of our method.

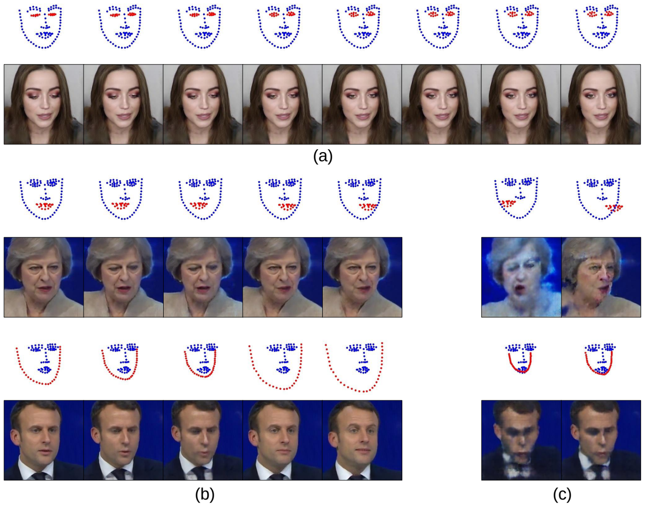

Fig. 4. Visual examples of face synthesis by editing landmarks.

3.3. Face Reenactment

In the second set of experiments, we build a face reenactment system based on our landmark to face synthesis method by adding a facial landmark extractor. Specifically, given an input face image, we extract facial landmarks, which are then fed to create a face preserving the facial expressions and orientations.

Compared methods. We compare with five state-of-the-art methods using input and target identities from the CelebV dataset, which are **GANimation** [18], **X2Face** [29], **FO**

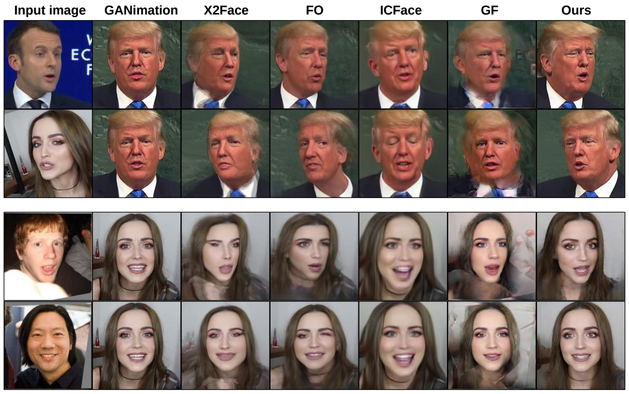

Fig. 5. Qualitative comparison of each method on CelebV dataset (left) and wild images (right). See text for details.

Table 1. Quantitative evaluations. See text for details.

Methods	LMK↓	SSIM↑	ID↓
GANimation	3.09	0.46	0.37
X2Face	1.14	0.61	0.45
FO	1.02	0.63	0.40
ICFace	3.33	0.47	0.47
GF	1.63	0.55	0.41
LandmarkGAN	**0.77**	**0.68**	**0.31**

[30], **ICFace** [19] and **GF** [31]. Specifically, GANimation achieves facial expression synthesis based on Action Units (AU) annotations from a single image[2]. X2Face is a self-supervised network that can transfer the pose and expression of a source face to a target face. Similar to GANimation, FO also achieves the face animation by decoupling the appearance and motion information using a self-supervised formulation. ICFace achieves face reenactment using human interpretable control signals such as head pose angles and AU values. GF is designed for pose-guided person image generation using global-flow local-attention model.

Visual comparison. Fig.5 shows qualitative comparison of each method on CelebV dataset (top) and wild images (bottom), which shows our method is better at preserving facial expressions and head orientations. Note for disentanglement based methods, we follow the instructions of each method to select the reference images of target identity and utilize the input images as driven images.

Quantitative evaluations. We next compare results of face reenactment quantitatively using three metrics: *landmark difference (LMK)*, *SSIM*, *identity difference between synthesized and target face (ID)*. LMK aims to evaluate the fidelity of the synthesized face images in terms of preserving the landmarks of the input image. Since the identity is changed during face reenactment, directly evaluating the landmarks preserving between the source and target identity is difficult due to the large variety of face shape in different identity. To this end, we use the synthesized target face of each method as input again

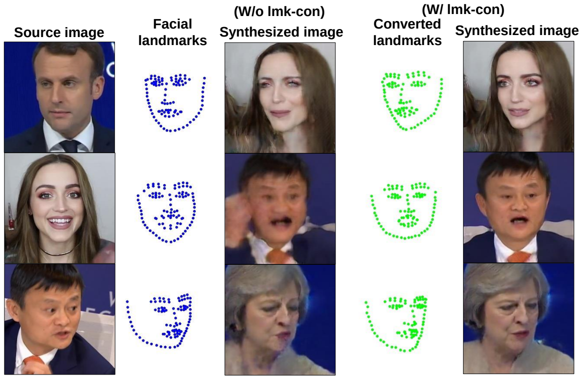

Fig. 6. Ablation study on the effectiveness of landmark converter.

to synthesize the source face. In this way, we first normalize the landmarks to $[0, 1]$ and then calculate the ℓ_2 landmark difference between the original source face and synthesized source face. SSIM [32] is used in recent works [13, 33, 12] to evaluate visual quality. Since the ground truth of synthesized target face is not existed in our experiment and it is not appropriate to directly calculate the SSIM score of synthesized target face referred on the source face, we use the similar setting as in LMK. We then evaluate the quality of identity swap. Specifically, we select a frontal face of the target identity from CelebV dataset as reference for each method, then we calculate the face recognition score (*e.g.*, Dlib [34]) between the reference and synthesized face of target. Table 1 shows the details of quantitative evaluation, which demonstrates the effectiveness of our method.

3.4. Ablation Study of Landmark Converter

To demonstrate the role of landmark converter, we conduct a set of experiments by directly feeding the extracted facial landmarks without using landmark converter. Fig.6 shows the results without (`w/o lmk-con`) and with landmark converter (`w/ lmk-con`). The quality of synthesized faces is significantly degraded without using the landmark converter, exhibited as blurring and inaccurate face shapes. The fourth and fifth column are the landmarks after landmark converter and corresponding synthesized results. It reveals the output of landmark converter has an intuitive deformation, which adjusts landmarks to better fit the target identity.

4. CONCLUSIONS

In this work, we describe a new method, known as LandmarkGAN, to synthesize faces from facial landmarks. Facial landmarks are a natural, intuitive, and effective representation for facial expressions and orientations, which are independent from the target's texture or color and background scene. Our model consists of two components: a landmark converter which converts the input facial landmarks to those of target face, and a target-specific landmark-to-face generator which synthesize a target face based on converted facial landmarks. Face synthesis and reenactment experiments conducted on CelebV dataset demonstrate the effectiveness of our method.

[2] Note the input of GANimation is the central face instead of the whole face. For comparison with others, we paste the synthesized face area back to the same location in original image.

5. REFERENCES

[1] Emily L Denton, Soumith Chintala, Rob Fergus, et al., "Deep generative image models using a laplacian pyramid of adversarial networks," in *NeurIPS*, 2015.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[3] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[6] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *CVPR*, 2018.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018.

[8] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, 2018.

[9] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy, "Reenactgan: Learning to reenact faces via boundary transfer," in *ECCV*, 2018.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[11] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[12] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *CVPR*, 2019.

[13] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *AAAI*, 2020.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[16] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," *arXiv preprint arXiv:1902.09212*, 2019.

[17] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *ICCV*, 2019.

[18] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *ECCV*, 2018.

[19] Soumya Tripathy, Juho Kannala, and Esa Rahtu, "Icface: Interpretable and controllable face reenactment using gans," in *WACV*, 2020.

[20] Xing Di, Vishwanath A Sindagi, and Vishal M Patel, "Gp-gan: Gender preserving gan for synthesizing faces from landmarks," in *ICPR*, 2018.

[21] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016.

[23] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast, "Numerical coordinate regression with convolutional neural networks," in *ECCV*, 2018.

[24] Diogo C Luvizon, Hedi Tabia, and David Picard, "Human pose regression by combining indirect part detection and contextual information," *arXiv preprint arXiv:1710.02322*, 2017.

[25] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz, "Few-shot unsupervised image-to-image translation," *arXiv preprint arXiv:1905.01723*, 2019.

[26] Geoffrey Hinton, "Neural networks for machine learning - lecture 6a - overview of mini-batch gradient descent.," 2012.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.

[28] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *CVPR*, 2018.

[29] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *ECCV*, 2018.

[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First order motion model for image animation," in *NeurIPS*, 2019.

[31] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li, "Deep image spatial transformation for person image generation," in *CVPR*, 2020.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[33] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan, "Freenet: Multi-identity face reenactment," in *CVPR*, 2020.

[34] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, 2009.