Face Image Animation with Adversarial Learning and Motion Transfer[1]

https://doi.org/10.3991/ijim.v16i10.30047

Abdulamir A. Karim(✉), Suha Mohammed Saleh
Computer Science Department, Technology University, Baghdad, Iraq
abdulamir.a.karim@uotechnology.edu.iq

Abstract—Significant advances have been made in facial image animation from a single image. Nonetheless, generating convincing facial feature movements remains a complex challenge in computer graphics. The purpose of this study is to develop an efficient and effective approach for transferring motion from a source video to a single facial image by governing the position and expression of the face in the video to generate a new video imitating the source image. Compared to prior methods that focus solely on manipulating facial expressions, this model has been trained to distinguish the moving foreground from the background image and to create motions such as facial rotation and translation as well as small local motions such as gaze shift. The proposed technique uses generative adversarial networks GANs with a motion transfer model. The network forecasts photorealistic video frames for a given target image using synthetic input in renderings from a parametric face model. The authenticity in this postprocessing conversion is attained by precise image manipulation. Thorough adversarial training is used to produce greater accuracy in this postprocessing conversion. Although more improvements to face landmark identification on videos and face super-resolution techniques have been made to improve the results, the proposed technique can provide more coherent videos with improved visual quality, resulting in more aligned landmark sequences for training. In addition, experiments indicate that we obtain superior results compared to those obtained by the state-of-the-art image-driven technique with PSNR 30.74 and SSIM 0.90.

Keywords—adversarial learning, face image super-resolution, image-to-video, motion transfer

1 Introduction

Facial image animation is the process of moving a single image to a target video so that it smoothly substitutes an existing face in the target while maintaining a realistic appearance. Face transform (also known as re-enactment) is a technique in which a video's facial movements and deformations are used to influence the motions and

[1] The authors hereby confirm that they have obtained the consent of the persons depicted in the photographs for publication.

deformations in another video or image. Face transform has recently been the focus of researchers because of its possible applications in video games, video editing, news broadcasting, virtual reality, and elsewhere [1–3]. The modeling and display of virtual face images are a persistent, difficult challenge in graphic design [4]. Due to the complexity of deformation, material characteristics, and phenomena such as occlusion and dis-occlusion, the task of developing efficient representations capable of creating high-quality renderings is a complex one. These issues are prevalent in frameworks that attempt to include the eyes and mouth cavity in their entirety. A geometry-based method would need precise modeling of the eyes and eyelids as well as the tongue, teeth, and gums, all of which are prone to artifacts, have poor visual quality, and require substantial manual involvement. Previous research suggested exchanging or re-enacting procedures to transmit or manipulate facial appearances and depended on underlying 3D face models [5]. Face forms were approximated or fixed based on the supplied image [6, 7]. The input photos were then aligned with the 3D geometry and subsequently utilized for swapping or re-enactment [8]. Researchers have developed techniques for the autonomous synthesis and augmentation of visual data during the last few years. Numerous techniques evolved from generative adversarial networks (GANs) [9] and variational autoencoders (VAEs) [10]. Additional data, such as conditioning labels, are included in these systems (e.g. indicating body poses and facial expressions). They are entirely data-driven, depending on a vast quantity of training data to find a latent representation of the visual inputs needed for the image generation [11–14]. Multiple face re-enactment approaches were motivated by using conditional generative adversarial networks (CGANs) [15] to transfer an image depicting actual data from one domain to another [16]. Siarohin et al. [17] have proposed a technique for image animation in two steps. First, dense motion is estimated using a first-order motion model; the image is then refined using an image generator. This method is entirely self-supervised and significantly improves the quality of animated face videos. The identity of the source face is retained, and the facial movements are detailed and constant in time. Despite the success of this method, face deformation and errors were present in the generated videos, as shown in Figure 1. Inspired by the method suggested in [17], our method looks at how to create a video of a target face (using only one image of the target) that replicates the behaviors of a source face when the source and target faces are different entities. More precisely, this method enables a source face to control the target's facial expressions, eye movement, rigid head position, and, to some extent, the target's face identity. These dimensions can be tweaked independently or in tandem with the others. The entire head and hair automatically create the target frames, a natural upper body, and a background that adheres to the altered head. The following summarizes the significant contributions of this study:

- A ResNet is used for the encoder section of the motion transfer model to enhance the motion transformation from the source video to the target image.
- The Viola-Jones face detection algorithm is used to improve landmark detection, which significantly reduces face deformation.
- A novel face super-resolution model is used to improve the output video quality.
- A hybrid strategy, SWATS (switch adaptive to stochastic), trains the model with an adaptive method such as the Adam optimizer and switches to a stochastic method such as stochastic gradient descent (SGD) to improve the learning process.

- Our method achieves better results than state-of-the-art methods, as discussed in the results and discussion section.

The rest of the article is structured as follows: Section 2 summarizes significant research on facial animation. Section 3 details the method used in this study. Section 4 details the experiment. Section 5 discusses the research and results. Finally, in Section 6, we conclude.

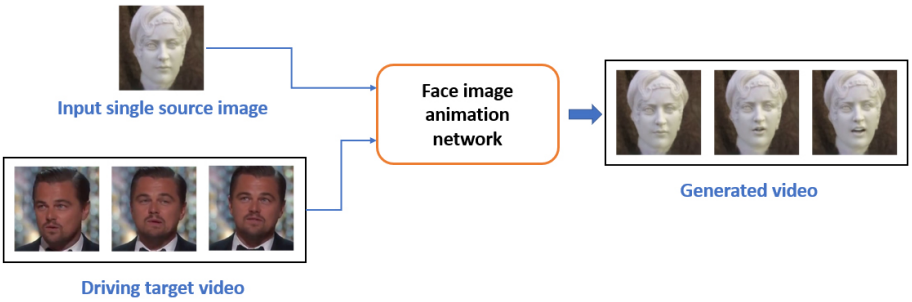

Fig. 1. Overview of the suggested method for generating video from a single source image controlled by driving frames from target video

2 Related work

Controlling the animation of facial images has an approximately two-decade history. Although these methods were first created to resolve privacy concerns [18], they are becoming more popular for amusement [19] and entertainment [20].

2.1 Face animation based on 3D methods

Historically, face animation (or puppeteering) was performed by fitting a 3D morphable model to a single image and manipulating the anticipated parameters [21]. Subsequent work expanded on the fitting of 3D morphable by including high-level features [22, 23], adding more images [24] or 3D scans [25], or immediately learning 3D morphable model parameters from RGB data without the need for labeled data. Unfortunately, although these approaches are quite precise, they are highly domain-specific and their performance suffers dramatically in challenging circumstances, such as when occlusions are present [26].

2.2 GANs and VAE based methods

It has been demonstrated that GANs [9] generate fake images with the same distribution as the target domain. Therefore, VGAN [27], a 3D convolutional GAN capable of concurrently generating all target video frames, was introduced by Vondrick et al. Using the same method, TGAN is a GAN-based model created by Saito et al. [28] that can create multiple frames simultaneously. The visual quality of these processes' outputs,

on the other hand, is usually poor. Recurrent neural networks trained in an adversarial approach have been used in recent video production processes. Wang et al. [14], for example, introduced the Conditional MultiMode Network (CMM-Net), a deep architecture that combines a conditional long short-term memory (LSTM) network and a VAE to create face videos. In addition, MoCoGAN [29], a deep architecture based on an adversarial learning-trained recurrent neural network, was introduced by Tulyakov et al. Using conditional information such as category labels or static photographs as input, these algorithms can produce high-quality video frames representing desired activities. Lombardi et al. [30] and Slossberg et al. [31] proposed an alternative method. Instead of working in pixel space, they train a deep neural network using a 3D model with texture using a 128-dimensional vector. They use a high-quality face capture system comprising 40 synchronized machine vision cameras to reconstruct consistent 3D face geometry for each frame. Finally, they train a VAE to synthesize geometry as well as a view-dependent texture using 3D geometry and the average texture, which facilitates the portrayal of high-quality 3D face sequences. Additionally, they demonstrated a technique for modifying face expressions by imposing position constraints on vertices, allowing greater creative control over facial emotions.

2.3 Controlling image production with a multi-modal method

[32] use several samples of the source face to create an identity embedder, then combine the rasterized landmarks image and the identity embedded vector in an image generator to create the target image. By including high-frequency characteristics in their later work [33], they improved the speed of neural rendering without losing visual quality by combining the optical flow warped version of the input image with the synthesized intermediate images. On a per-frame basis, [34] handled the problem of motion transfer by framing it inside an image-to-image translation paradigm. Furthermore, it suggested imposing geographical and temporal constraints. In [35], they stressed the significance of video synthesis, particularly temporal dynamics. Finally, [36] presented X2Face, a deep architecture for turning a face image into motion patterns created by another face or modality, such as audio, from a face input image. They demonstrated that a data-driven system can animate still images of faces without using 3D rendering.

Robust facial landmark detection algorithms have been developed over time. Previous research has used Dlib [37] or FAN4 [38] to train their algorithms to detect face landmarks. However, these generate outputs that are riddled with temporal graphical anomalies. As a result, this study incorporates the Viola-Jones face detection approach. This method may result in a shift in focus, and reliable facial landmarks enhance its efficiency.

3 The proposed method

The objective is to make a face movie with a source face image and a sequence of driving face landmarks in which the face's motion matches that of the driving face landmarks. Face images are identified through a vector concatenating each landmark's coordinates. The task is made more accessible by creating simply one face image. A new face image I_t^n is created for each triple of the source face image I_s, source face

landmark vector I_{LM}, and driving landmark vector Dr_{lm}^{n}, n = 1,2,..., N are then concatenated to create a face video; N denotes the frames count.

3.1 Method overview

Figure 2 illustrates our strategy in further detail. Two main networks are used to build an end-to-end model to generate a new face video. The first leading network is the dense network, which is the essential part; its role is to transfer the motion from the driving video to the source image I_s This network used a U-Net model with ResNet pre-trained on the ImageNet dataset for the encoder section of the key point detector to get the key points from I_s and I_D^n (the image from the driven video). The Viola-Jones face detection algorithm is employed for improving key point selection in this step. A ten-filter convolutional layer is added to the previously described network to generate feature maps of size *batch* × *height* × *width* × 10 as the output. Then, for calculating the location of the key points, the maximum value of each feature map is estimated using the Softmax function; this function is applied along the spatial axes (height and width). Therefore, the final key points will have a shape of *batch* × 10 × 2, where each keypoint has two values for X and Y coordinates. Then, the dense motion network takes the source image I_s, its keypoints I_{LM}, and the key points of the driving image Dr_{lm}^{n} to generate a dense motion map F^n, which denotes the per-pixel mapping. Next, bi-linear warping is applied to warp I_s through F^n to achieve an initial estimation E_I^n as in the following equation:

$$E_I^n = \mathrm{B}_w^b(I_s, F^n) \tag{1}$$

Then, the result of concatenating E_I^n and F^n is used as the input of the second leading network (the generator, which is a GAN network) to generate and refine the occluded regions. The generator generates the final image by using Eq. (2) as follows:

$$I_t^n = E_I^n \odot (1 - M^n) + R_I^n \odot M^n \tag{2}$$

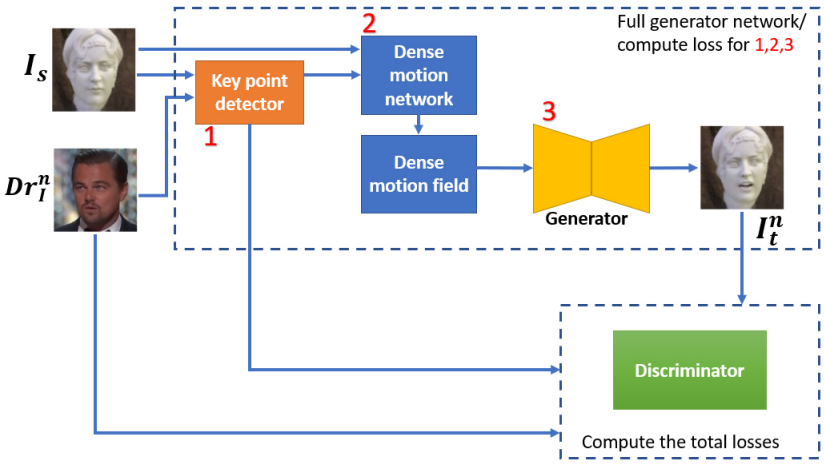

Fig. 2. The block diagram of the suggested network

3.2 The super-resolution model

A face super-resolution model is built to increase the quality of the generated video. This model comprises a series of convolutional layers, each with a 3 × 3 kernel, that perform downsampling operations, followed by a chain of deconvolutional layers that perform upsampling operations to process information at various spatial scales, as shown in Figure 3. This approach facilitates more precise mapping of the low-resolution LR and high-resolution HR faces by sharing low-level characteristics throughout the network. Five blocks of downsampling and five blocks of upsampling compose the model. In addition, a skip connection is inserted between layer i in the downsampling chain and layer $n\text{-}1 + i$ in the upsampling chain, where n is the total number of layers, to enhance pixel localization between LR and HR images. Using these skip connections, all image channels are concatenated.

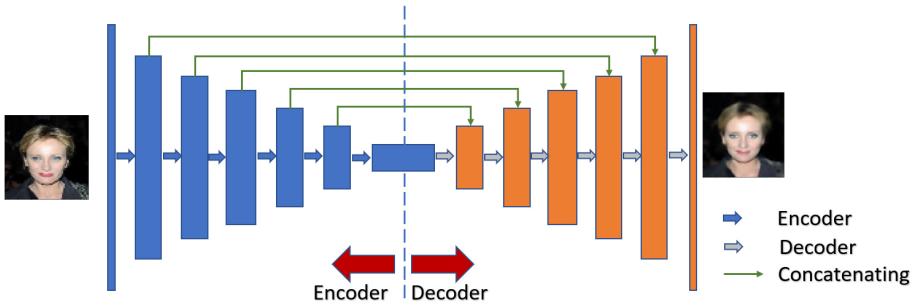

Fig. 3. The architecture of the proposed face super-resolution model

3.3 Training losses

In a self-supervised environment, the network is trained using a collection of face recordings. It detects face landmarks in submitted videos prior to training. The model picks pairs of the source and driving images from each training video at random throughout the training phase. Using the source image I_s, the source face landmarks vector I_{LM}, and the driving face landmarks vector Dr^n_{lm}, the network is trained to rebuild I^n_t from scratch. The perceptual loss is estimated as in Eq. (3):

$$L_p(Dr^n_I, I^n_t) = \sum_{j=1}^{j} | K_j(Dr^n_I) - K_j(I^n_t) | \tag{3}$$

where Dr^n_I is the image from the driven video, K_j is the jth channel feature extracted using VGG-19, and J is the feature channels count.

For the discriminator D part of the GAN network, least-square error loss is used as follows in Eq. (4):

$$L_D(Dr^n_I, I^n_t) = \left\| 1 - D(Dr^n_I) \right\| + \left\| D(I^n_t) \right\| \tag{4}$$

For the generator *G* part as well, the least-square error loss is defined as follows in Eq. (5):

$$L_G(Dr_I^n, I_t^n) = \left\| 1 - G(I_t^n) \right\| \tag{5}$$

In Eq. (6), the total of all prior losses is utilized as an objective function to be minimized:

$$L_{total} = L_p + L_D + L_G \tag{6}$$

Whereas for the face image super-resolution model, the loss function is calculated as follows in Eq. (7):

$$L_{SR}(HR_i - \widehat{HR}_i) = \frac{1}{N} \left\| HR_i - \widehat{HR}_i \right\| \tag{7}$$

where HR_i and $\widehat{HR}_i$ are the original image and the generated image by the model, respectively. The overall training procedure before applying the super-resolution model on the generated image is illustrated in Algorithm 1.

Algorithm 1: Training procedure

Input: Batch size B, I_s and Dr_I^n, n = 1,2,..., N, the maximum number of epochs E.
Output: Trained model
1: while Iteration < E do:
2: Use key points detector to estimate I_{LM} and Dr_{lm}^n.
3: Put I_s, I_{LM}, and Dr_{lm}^n into the dense network to generate a dense motion map F^n.
4: Calculate the initial estimation E_I^n using Eq. (1).
5: Use the generator network to generate the final image I_t^n according to Eq. (2).
6: Calculate the loss using Eq. (6).
7: Update the weights
8: **end** while

4 Experiments

The experiments used the VoxCeleb dataset of 22,496 face videos collected from YouTube. An initial bounding box is produced from the first video frame for pre-processing. This face is tracked until it deviates too far from its original place. The video frames are then cropped using the smallest crop that encompasses all bounding boxes. The method is continued until the sequence is completed. Finally, sequences with a resolution of less than 256 × 256 are removed; the remaining videos are enlarged to 256 × 256 while maintaining the aspect ratio.

4.1 Performance metrics

Two commonly used metrics are used to evaluate the method performance and compare them to state-of-the-art works:

- PSNR (peak signal-to-noise ratio) measured in *dB*, as defined in Eq. (8):

$$PSNR(Dr_I^n, I_t^n) = 20\log\left(\frac{L}{\sqrt{MSE(Dr_I^n, I_t^n)}}\right) \tag{8}$$

where L represents the maximum possible pixel value of the image, *MSE* is the mean squared error between the ground truth image and the image produced by the model. Thus, *PSNR* is in the range of (0, ∞], where a higher value is better.

- SSIM (structural similarity index measure) given by [39] is defined in Eq. (9):

$$SSIM(Dr_I^n, I_t^n) = \frac{1}{P}\sum_{i=1}^{P} sim(Dr_I^n, I_t^n), \tag{9}$$

where *sim* (.,.) is the similarity function that measures the structural similarity between the P image patches (Dr_I^n, I_t^n); This function is defined in Eq. (10) below:

$$sim(Dr_I^n, I_t^n) = \frac{(2u_1u_2 + C1)(2s_{12} + C2)}{(u_1^2 + u_2^2 + C1)(s_1^2 + s_2^2 + C2)}, \tag{10}$$

where u_1 and u_2 are the means of the local patches Dr_I^n and I_t^n, s_1^2 and s_2^2 denote their local variances, and S_{12} denotes the local covariances of Dr_I^n and I_t^n. C_1 and C_2 are constant hyperparameters. SSIM is in the range of (0, 1), where 1 indicates that the two images are identical.

4.2 Model training

A hybrid strategy, SWATS (switch adaptive to stochastic), trains the model with an adaptive method (in this experiment, Adam optimizer is used) and switches to a stochastic method (stochastic gradient descent, SGD) to improve the learning process. These experiments are conducted on Intel(R) Core (TM) i7-10750H CPU @ 2.60 GHz with 2.59 GHz, 32.0 GB of RAM, and a display adapter from NVIDIA GeForce GTX 1650 Ti.

5 Results and discussion

Extensive quantitative and qualitative experiments have been performed to verify the effectiveness of the suggested method, which was compared with the following state-of-the-art methods: X2Face (2018) [36], MoCoGAN (2018) [29], Monkey-Net (2019) [40], Few-shot (2019) [32], and PuppeteerGAN (2020) [41]. The results show that our suggested method performs better than all the above-mentioned methods. This approach is better than [17] at generating images with delicate local motion and controlling facial pose, whereas other methods struggle. As shown in Figure 4, this method can handle expressions even in large-angle difference and driving video with no glasses for a source image with glasses. Different positions and expressions, facial shapes, and hair occlusions are represented in certain samples. Quantitative comparison

results are shown in Table 1. PSNR and SSIM are computed for each method, and our proposed method achieved the highest values.

Table 1. Quantitative comparison with similar methods

Method	PSNR	SSIM
X2Face	22.54	0.72
MoCoGAN	17.36	0.54
MonkeyNet	30.87	0.74
Few-shot	N/A	0.72
PuppeteerGAN	N/A	0.73
The suggested method	**30.74**	**0.90**

Fig. 4. Samples of face animation results were produced by our suggested network

Furthermore, a selection of animation character images have been applied to these experiments and achieved good results, despite the relative paucity of facial details in these images. In real life, challenge images are made by collecting photographs of public figures from the Internet and generating face images using our suggested model trained on VoxCeleb1. Additionally, the testing is conducted across many identities, with the movement of one person's facial image influencing the movement of another person's facial image. Finally, a comparison is made with [17] (whose model is also based on VoxCeleb1). The visualization findings in Figure 5 demonstrate that our technique produces more realistic images with more strong gaze transitions and less form distortion. On the other hand, the visuals of [17] cannot track the direction of the gaze in driving images.

Fig. 5. Qualitative comparison with the state-of-the-art first-order method

Although this method may yield very accurate re-enactment results in many applications and conditions, it does have certain limits. Like many others of its kind, this strategy performs wonderfully within the constraints of the training corpus. However, extreme target head positions, such as extreme rotations or emotions, may degrade the video portrait's visual quality. Furthermore, because a parametric model is trained only on the face, the motions of the body, hair, or backdrop cannot be actively adjusted. Instead, the network eventually extrapolates for a particular head posture and calculates a realistic and consistent upper. This limitation can be resolved by training the model on the body and developing an enormous collection of conditioning photos based on the underlying body model. Finally, rather than constituting a restriction, the democratization of advanced high-quality video editing abilities afforded by this and other techniques calls for increased attention to be paid to verifying film authenticity, such as by invisible watermarking.The suggested technique does not require a significant amount of time-consuming, subject-specific data collecting and model training, thereby making face animation easier to achieve for non-experts.

6 Conclusion

A new approach for animating facial images has been demonstrated. The dense motion generation technique is used, followed by image generation. In addition, to improve the outcomes, a super-resolution model is applied. This approach may create both global (such as face rotation and translation) and fine local motion (such as gaze

change). Face landmark identification may also be accomplished using the Viola–Jones method, which significantly increases the visual quality and temporal coherence of created videos. Experiments have revealed that this approach is better than existing state-of-the-art image-driven algorithms in terms of outcomes. It is recommended that future work use the audio modality to significantly improve the quality of facial images produced.

7 References

[1] H. Kim et al., "Deep video portraits," ACM Transactions on Graphics, vol. 37, no. 4, 2018. https://doi.org/10.1145/3197517.3201283

[2] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7183–7192. https://doi.org/10.1109/ICCV.2019.00728

[3] R. L. Smith, P. Dasari, C. Lindsay, M. King, and K. Wells, "Dense motion propagation from sparse samples," Physics in Medicine and Biology, vol. 64, no. 20, 2019. https://doi.org/10.1088/1361-6560/ab41a0

[4] W. Paier, A. Hilsmann, and P. Eisert, "Interactive facial animation with deep neural networks," IET Computer Vision, vol. 14, no. 6, 2020. https://doi.org/10.1049/iet-cvi.2019.0790

[5] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, "Extreme 3D face reconstruction: Looking past occlusions," CVPR, 2018. https://doi.org/10.1109/CVPR.2018.00414

[6] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," Communications of the ACM, vol. 62, no. 1, 2019. https://doi.org/10.1145/3292039

[7] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," in ACM Transactions on Graphics, 2017, vol. 36, no. 4. https://doi.org/10.1145/3072959.3073640

[8] H. T. Salim, and I. A. Aljazaery, "Encryption of color image based on DNA strand and exponential factor," International journal of online and biomedical engineering (iJOE), vol. 18, no. 3, 2022. https://doi.org/10.3991/ijoe.v18i03.28021

[9] I. J. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2014, vol. 3, no. January. https://doi.org/10.3156/jsoft.29.5_177_2

[10] H. Tauma, "Enhanced Data Security of Communication System using Combined Encryption and Steganography," International Journal of Interactive Mobile Technologies, vol. 15, no. 16, pp. 144–157, 2021. https://doi.org/10.3991/ijim.v15i16.24557

[11] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," 2018. https://doi.org/10.1145/3240508.3240704

[12] Z. Geng, C. Cao, and S. Tulyakov, "3D guided fine-grained face manipulation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 9813–9822. https://doi.org/10.1109/CVPR.2019.01005

[13] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," 2018. https://doi.org/10.1109/CVPR.2018.00359

[14] H. Salim, and N. Alseelawi, "A novel method of multimodal medical image fusion based on hybrid approach of NSCT and DTCWT," International journal of online and biomedical engineering, vol. 18, no. 3, 2022. https://doi.org/10.3991/ijoe.v18i03.28011

[15] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[16] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11214 LNCS. https://doi.org/10.1007/978-3-030-01249-6_50

[17] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in Advances in Neural Information Processing Systems, 2019, vol. 32.

[18] S. Mosaddegh, L. Simon, and F. Jurie, "Photorealistic face de-identification by aggregating donors' face components," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9005. https://doi.org/10.1007/978-3-319-16811-1_11

[19] I. Kemelmacher-Shlizerman, "Transfiguring portraits," in ACM Transactions on Graphics, 2016, vol. 35, no. 4. https://doi.org/10.1145/2897824.2925871

[20] L. Wolf, Z. Freund, and S. Avidan, "An eye for an eye: A single camera gaze-replacement method," 2010. https://doi.org/10.1109/CVPR.2010.5540133

[21] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," 1999. https://doi.org/10.1145/311535.311556

[22] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, "Extreme 3D face reconstruction: Seeing through occlusions," 2018. https://doi.org/10.1109/CVPR.2018.00414

[23] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 2326–2335. https://doi.org/10.1109/CVPR.2017.250

[24] J. Roth, Y. Tong, and X. Liu, "Adaptive 3D face reconstruction from unconstrained photo collections," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, 2017. https://doi.org/10.1109/TPAMI.2016.2636829

[25] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3D morphable models," International Journal of Computer Vision, vol. 126, no. 2–4, 2018. https://doi.org/10.1007/s11263-017-1009-7

[26] M. Zollhöfer et al., "State of the art on monocular 3D face reconstruction, tracking, and applications," Computer Graphics Forum, vol. 37, no. 2, 2018. https://doi.org/10.1111/cgf.13382

[27] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," 2016.

[28] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2849–2858. https://doi.org/10.1109/ICCV.2017.308

[29] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," 2018. https://doi.org/10.1109/CVPR.2018.00165

[30] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," ACM Transactions on Graphics, vol. 37, no. 4, 2018. https://doi.org/10.1145/3197517.3201401

[31] R. Slossberg, G. Shamai, and R. Kimmel, "High quality facial surface and texture synthesis via generative adversarial networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11131 LNCS. https://doi.org/10.1007/978-3-030-11015-4_36

[32] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9458–9467. https://doi.org/10.1109/ICCV.2019.00955

[33] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, "Fast Bi-Layer Neural Synthesis of One-Shot Realistic Head Avatars," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 12357 LNCS. https://doi.org/10.1007/978-3-030-58610-2_31

[34] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5932–5941. https://doi.org/10.1109/ICCV.2019.00603

[35] T. C. Wang et al., "Video-to-video synthesis," in Advances in Neural Information Processing Systems, 2018.

[36] O. Wiles, A. S. Koepke, and A. Zisserman, "X2Face: A network for controlling face generation using images, audio, and pose codes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11217 LNCS. https://doi.org/10.1007/978-3-030-01261-8_41

[37] D. E. King, "Dlib-ml: A machine learning toolkit," Journal of Machine Learning Research, vol. 10, 2009.

[38] A. Bulat, G. Tzimiropoulos, and U. Kingdom, "How far are we from solving the 2D & 3D Face Alignment problem?," ICCV, 2017. https://doi.org/10.1109/ICCV.2017.116

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, 2004, https://doi.org/10.1109/TIP.2003.819861

[40] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 2372–2381. https://doi.org/10.1109/CVPR.2019.00248

[41] Z. Chen, C. Wang, B. Yuan, and D. Tao, "PuppeteerGAN: Arbitrary portrait animation with semantic-aware appearance transformation," 2020. https://doi.org/10.1109/CVPR42600.2020.01353

8 Authors

Abdulamir A. Karim Prof. Dr. Computer Science Department, Technology University, Bagdad, Iraq. E-mail: abdulamir.a.karim@uotechnology.edu.iq

Suha Mohammed Ph.D student in Computer Science Department, Technology University, Baghdad, Iraq. E-mail: cs.19.83@grad.uotechnology.edu.iq

Article submitted 2022-02-03. Resubmitted 2022-03-21. Final acceptance 2022-03-23. Final version published as submitted by the authors.