DFT-NET: DISENTANGLMENT OF FACE DEFORMATION AND TEXTURE SYNTHESIS FOR EXPRESSION EDITING

Jinghui Wang[⋆†] *Jie Zhang*[†] *Zijia Lu*[†] *Shiguang Shan*[†]

[⋆]Beijing University of Posts and Telecommunications
[†]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

ABSTRACT

This paper presents a novel deep architecture DFT-Net that combines the advantages of Generative Adversarial Networks (GANs) and warp mechanism for expression editing. Recent generative models leverage Action Units as annotations and show more flexible expression manipulation than previous approaches using other guiding information. However, those methods bring inevitable artifacts where facial components deform (e.g. eyes from open to close), for the structural defect in modeling shape variations without geometric guidance such as facial landmarks. Our approach explicitly disentangles face deformations and appearance details by constructing two parallel networks, one that learns an appearance flow for 2D warps and the other generates corresponding texture and hallucinates hidden regions such as mouth interior. Experimental results show our method outperforms the state-of-the-art on various expression editing tasks.

Index Terms— Expression Editing, GANs, Action Units, Flow Fields

1. INTRODUCTION AND RELATED WORKS

Over the years, a number of works have been devoted to expression editing or face animation in the field of computer graphics, computer vision and pattern recognition. However, synthesizing high quality facial expression images is still challenging because facial expression transformations are highly non-linear.

Most of the computer graphics techniques resort to finding correspondences between source facial features and target images [1]. The correspondences are then utilized to warp face images to the desired expression [2–5]. While these methods can generate realistic images with high resolution, they cannot transfer dynamic textures, such as self-shadowing in wrinkles and creases, or hallucinate hidden regions like mouth interior. With the development of Generative Adversarial Networks, several approaches based on generative models have been proposed. StarGAN [6] is able to edit other facial attributes besides expressions such as gender, hair color or age. However, it can only change expressions among discrete number of expressions. [7,8] utilize facial landmarks as global geometric guidance to control the process of expression synthesis, which enable continuous manipulation of expressions and show that such geometric cues could significantly boost the performance. Despite their improvements, facial landmarks are essentially not enough to represent fine-scale details such as wrinkles and creases. Instead, [9, 10] leverage activated Action Units (AUs) [11] to anatomically control the generation process. Nevertheless, close inspections to the generated images usually reveal blurriness or artifacts.

Since synthesizing faces by hallucinating RGB values from scratch is very difficult, recently some works have been proposed to combine the advantages of graphics techniques and generative models. [12] hybridizes variational autoencoder [13] and the classical flow-based approach [4, 14] to get higher perceptual quality. However, it cannot inpaint unseen areas in the source image and often falls short in achieving the realism at the shading induced by wrinkles or creases. For animating a source portrait with a driving video, [15] build a two-stages framework which first captures the landmark correspondence between the source image and the driving video for light-weighted 2D warping and then appends two GANs to synthesize fine-scale details and mouth interiors. Although it can generate compelling perceptual results, it requires both expressions of the initial frame from the driving video and the source image to be neutral.

To address those limitations, we introduce a purely data-driving novel deep architecture, DFT-Net, conditioned on AUs for image-to-image expression editing. Our method explicitly factorizes the generating process into face deformations and texture hallucinations. Specifically, we build a two-path network, one path performs 2D warping by predicting an appearance flow [16] to align the input image to the target expression; the other regresses a texture mask which contains fine-scale textures and hallucinated hidden areas. Then we leverage attention mechanism to combine the warped image and the texture mask for the final synthesis. Different from GANimation [9] where the generated attention mask is used to guide the weighed addition between the source face image and the generated texture mask, our model predicts the attention mask to indicate the weights of the aligned face and the texture mask. Experiments demonstrate that DFT-Net can synthesis more photorealistic images than previous generative approaches.

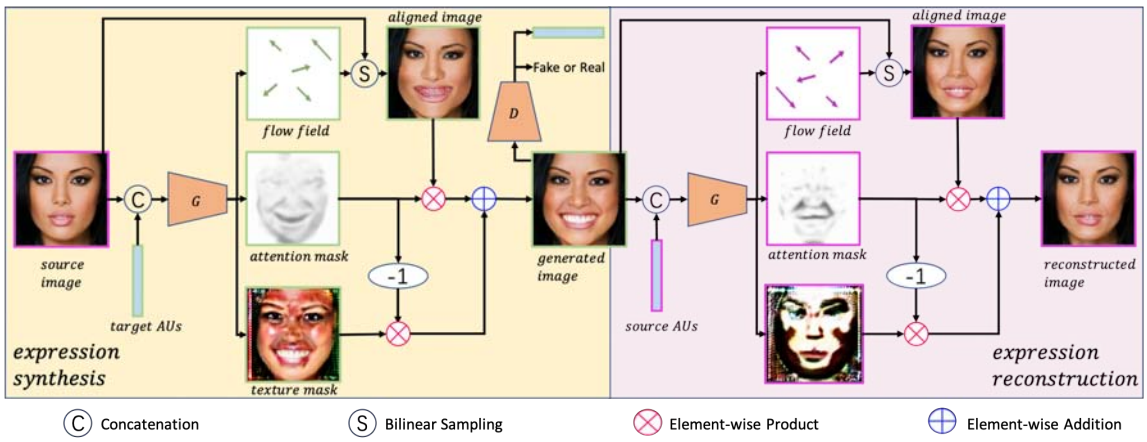

Fig. 1: The proposed DFT-Net framework. DFT-Net consists of a generator G and a discriminator D. G outputs flow fields for face deformations, attention masks and texture masks. D is used to judge the reality of the generated images and predict their AUs. The generated images are fed back to G to reconstruct the source images.

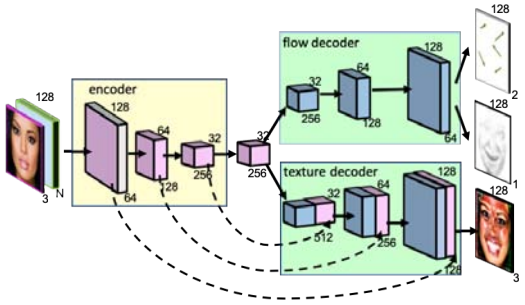

Fig. 2: The architecture of G. Each action unit is expanded as the same height and width as the input image.

2. PROPOSED MODEL

As illustrated in Fig. 1 , the proposed DFT-Net consists of a generator G and a discriminator D. G consists of an encoder and two decoders: the flow decoder F and the texture decoder T. The source image concatenated with AUs of the target expression is first encoded to a latent vector. Then the flow decoder generates a flow field map to align the input face to the desired shape by bilinear sampling. Parallelly, the texture decoder outputs a texture mask which contains fine-scale details and pixels of the hidden regions on the input (e.g. mouth interior). To determine to which extent each pixel of the texture mask will contribute to the final result, an attention mask is generated by the flow decoder along with the flow field map. Lastly, the generated image is fed into the same generator with the AUs of the source image to reconstruct the original expression, which enables unsupervised training, i.e., no pairs of the images of the same person under different expressions are needed.

2.1. Generator

The generator G generally adopts a variant structure of encoder-decoder, whereas the decoder is bifurcated. The motivation behind the design is that we found the original one-path generator in GANimation failed in some cases of geometric transformation. As shown in Fig. 3 , it can be clearly observed that the eyelids are generated by filling the color of the area of the eyeballs, rather than reuse the pixels around the eyes. Then we built a 2D warp model [16] and found it cannot carry the richness of facial expressions but was able to simulate facial deformations. As a result, we manage to combine their advantages by incorporating the warping manipulation into the generator. Besides, we find that adding skip connections across the encoder and the texture decoder works well for reducing the artifacts. The detailed structure of G is shown in Fig.2 . Different from [15] that warps the source image by graphic techniques to guide the following generation, we propose an end-to-end framework to integrally consider the two parts for joint training.

Specifically, we first encode the source image $X_s \in \mathbf{R}^{H\times W\times 3}$ and the target AUs $y_t \in [\mathbf{0}...\mathbf{1}]^{H\times W\times N}$ into a low-dimensional latent code c :

$$c = E(X_s, y_t) \ , \tag{1}$$

where E denotes the encoder. y_t is a N-channel image with the same size as the source image. Each channel represents a single Action Unit, where the value of each pixel is the same as the magnitude of the corresponding AU.

Secondly, we append two parallel decoders, flow decoder F and texture decoder T. The former predicts an appearance flow field [16] $\Phi \in \mathbf{R}^{H\times W\times 2}$ for warping and an attention mask $A \in \{0, ..., 1\}^{H\times W\times 1}$ to merge the aligned image X_a and the texture mask $B \in \mathbf{R}^{H\times W\times 3}$ generated by T :

$$\Phi, A = F(c) \ . \tag{2}$$

With Φ, the aligned image is generated by the bilinear sampling mechanism [17]. Note that the sampling operation is differentiable. Here we simplify the sampling function as f:

$$X_a = f(\Phi, X_s). \tag{3}$$

Fig. 3: Results of eyelid synthesis on Multi-PIE and EmotioNet. The leftmost column shows the source protraits. The remaining columns are the eyes cropped from source protraits, aligned images, our final results and GANimation's results.

The texture decoder decodes the latent vector c to a texture mask $B \in \mathbf{R}^{H\times W\times 3}$ which contains self-shadows, creases and hallucinated pixels of hidden areas:

$$B = T(c)\ . \tag{4}$$

Lastly the final image $\hat{X}_t \in \mathbf{R}^{H\times W\times 3}$ is governed by:

$$\hat{X}_t = (1-A)\cdot B + A\cdot X_a\ . \tag{5}$$

2.2. Discriminator

This network is trained to detect the generator's "fakes". We adopt the architecture from [18] that maps the source image or the generated final image to a matrix $M \in \mathbf{R}^{H/2^6\times W/2^6}$, which may ignore the subtle structures of local image patches. Similar to GANimation, we add an auxiliary regressor to estimate the AUs' magnitude $\hat{y}_t$ of the generated image.

2.3. Model Objective

2.3.1. Adversarial Loss

We use WGAN-GP [19] to stabilize the training process. Let P_s be the data distribution of the input image, $P_{\widetilde{X}}$ the random interpolation distribution, the loss can be formulated as

$$\begin{aligned}\mathcal{L}_{adv} = E_{X_s\sim\mathbb{P}_s}[D_A(G(X_s|y_t)) - D_A(X_s)] + \\ \lambda_{\text{gp}}\mathbb{E}_{\widetilde{X}\sim\mathbb{P}_{\widetilde{X}}}\left[(\|\nabla_{\widetilde{X}} D_A(\widetilde{X})\|_2 - 1)^2\right],\end{aligned} \tag{6}$$

where λ_{gp} denotes the penalty coefficient.

2.3.2. Conditional Loss

The discriminator also predicts the AUs' magnitude of the input images or the generated results. The conditional loss $\mathcal{L}_c$ enforces the generator to learn to satisfy the target facial expression:

$$\begin{aligned}\mathcal{L}_{cond} = E_{X_s\sim\mathbb{P}_s}[\|D_C(G(X_s|y_t)] - y_t\|_2^2] + \\ \mathbb{E}_{X_s\sim\mathbb{P}_s}[\|D_C(X_s) - y_s\|_2^2].\end{aligned} \tag{7}$$

Fig. 4: Details of the intermediate results. The aligned images, texture masks, attention masks, our final results and GANimation's results are exhibited from left to right, Darker regions in attention masks indicates those areas are more relevent to texture masks. Our method performs better than GANimation in terms of face artifacts.

2.3.3. Reconstruction Loss

To guarantee that the translated images preserve the identities of the input portraits. As in [20], we apply a cycle consistency loss to G, formulated as

$$\mathcal{L}_{rec} = E_{X_s\sim\mathbb{P}_s}[\|X_s - G(G(X_s|y_t)|y_s)\|_1], \tag{8}$$

where G takes the translated image $\hat{X}_t$ =$G(X_s|y_t)$ and source AUs y_s to reconstruct the original image X_s.

2.3.4. TV Loss

To facilitate the learning of the texture decoder T, it requires the flow field Φ to be spatially smooth. We perform the TV regularization on Φ:

$$\mathcal{L}_{tv} = E_{X_s\sim\mathbb{P}_s}\left[\sum_{i,j}^{H,W}(\Phi_{i+1,j} - \Phi_{i,j})^2 + (\Phi_{i,j+1} - \Phi_{i,j})^2\right]. \tag{9}$$

2.3.5. Attention Regularization Loss

As the flow field Φ can easily saturate to 0 which makes that $X_s = f(\Phi, X_s)$, that is, the flow decoder has no effect. Thus,

we impose a L1 regularization on the attention mask:

$$\mathcal{L}_{att} = -E_{X_s \sim \mathbb{P}_s}[\|A\|_1]. \quad (10)$$

3. EXPERIMENTS

In this section we first introduce our training settings and then show some experimental results. We compared our method with GANimation since it is currently the state-of-the-art method using AUs for expression editing to our best knowledge.

3.1. Training Settings

We use a subset of 200,000 samples of EmotioNet [21] dataset for training. The images are cropped and resized to 128×128, where the faces are centered. The continuous AUs annotation are extracted by the AU detector [22]. The target AUs are firstly sampled from the whole training set and then normalized between 0 and 1. During optimization, Adam [23] with learning rate of 0.0001, beta1 = 0.5, beta2 = 0.999 is adopted as the optimizer. We trained the model on a single GeForce GTX 1080 Ti GPU for 30 epochs with linearly decay over the last 10 epochs. Every 5 optimization steps of the discriminator we perform a single optimization step of the generator. The weight coefficients for the loss terms above are set to $\lambda_{\text{gp}} = 10$, $\lambda_{\text{adv}} = 1.0$,$\lambda_{\text{att}} = 1.0$, $\lambda_{\text{tv}} = 0.0001$, $\lambda_{\text{cond}} = 4000$, $\lambda_{\text{cyc}} = 10$. Our model is implemented using Pytorch v4.0, CUDNN v7.0, CUDA v9.0. The average inference times of GANimation and DFT-NET are 12.11ms and 7.32 ms per sample respectively.

3.2. Discrete Emotions Editing

We first compare our model's ability for discrete emotions editing against GANimation. The target AUs of each discrete emotion are extracted from RaFD dataset [24]. As it can be seen in Fig. 4 , our method can significantly reduce the blurriness around mouths and the checkerboard artifacts on the bridges of noses.

3.3. Eyelid synthesis

Our model significantly improves the quality of eyelid synthesis. By varying the specific AU (AU 45), we can controll the status of eyelids flexibly. Previous generative methods tend to simply add some color block to mask eyeballs, which lead to obvious artifacts that eyeballs are still visible.
DFT-Net tackles the problem through simultaneously warping faces and generating the target face appearance. As illustrated in Fig. 3 , pixels of eyelids are sampled around eyes to generate aligned image firstly. Based on the aligned image, we can get more realistic final results than GANimation.

3.4. Expression interpolation

For this experiment, giving a source image and a target image, we extract the AUs from the target image and increase the levels of their values (0.5, 0.75, 1.0) to gradually transfer the expression of the source image to the target. Examples are shown in Fig. 5 . We can see that DFT-Net generates much more cleaner results. For some cases that the input face has extreme expressions, generative models still struggle to handle them. It is mainly because of the lack of those "hard cases" in training data and the insufficient number of continuous AUs extracted by existing methods.

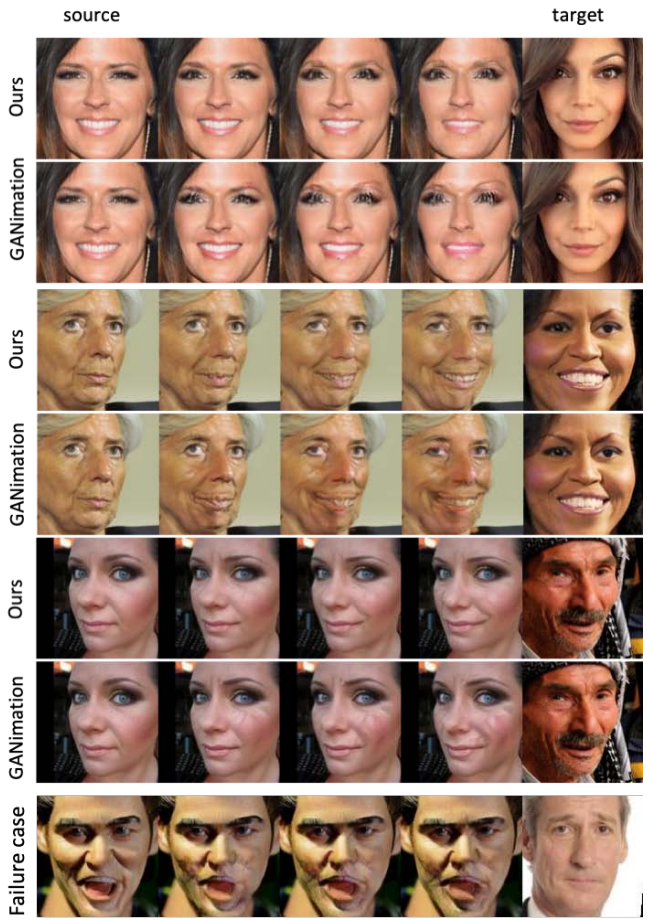

Fig. 5: Results of EmotioNet database for expression interpolation. Thee leftmost column shows the source portraits and the rightmost column shows the target portraits. From left to right, the remainders present the gradually transferred results. The last row shows one of the failure cases of DFT-NET.

4. CONCLUSION

We present a purely data-driven model, DFT-Net, trained in a fully unsupervised manner for expression editing. It combines 2D warps, GANs and attention mechanism that ensures both face deformations and fine-scale details synthesis. Experiments demonstrate that DFT-Net performs better than previous generative methods, especially for eyelid synthesis.

5. REFERENCES

[1] Barry-John Theobald, Iain Matthews, Michael Mangini, Jeffrey R Spies, Timothy R Brick, Jeffrey F Cohn, and Steven M Boker, "Mapping and manipulating facial expression," *Language and speech*, vol. 52, no. 2-3, pp. 369–386, 2009.

[2] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.

[3] Pia Breuer, Kwang-In Kim, Wolf Kienzle, Bernhard Scholkopf, and Volker Blanz, "Automatic 3d face reconstruction from single images or video," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–8.

[4] Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas, "Facial expression editing in video using a temporally-smooth factorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 861–868.

[5] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen, "Bringing portraits to life," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 196, 2017.

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[7] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang, "Geometry-contrastive generative adversarial network for facial expression synthesis," *arXiv preprint arXiv:1802.01822*, 2018.

[8] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan, "Geometry guided adversarial facial expression synthesis," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 627–635.

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

[10] Hai X Pham, Yuting Wang, and Vladimir Pavlovic, "Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network," *arXiv preprint arXiv:1803.07716*, 2018.

[11] E Friesen and P Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, 1978.

[12] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala, "Semantic facial expression editing using autoencoded flow," *arXiv preprint arXiv:1611.09961*, 2016.

[13] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[14] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas, "Expression flow for 3d-aware face component transfer," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, pp. 60, 2011.

[15] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou, "Warp-guided gans for single-photo facial animation," *ACM Transactions on Graphics*, vol. 37, no. 6, 2018.

[16] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

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

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

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

[21] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.

[22] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 6, pp. 1–6.

[23] D Kinga and J Ba Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015, vol. 5.

[24] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.