2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10223136

MULTI TASK-BASED FACIAL EXPRESSION SYNTHESIS WITH SUPERVISION LEARNING AND FEATURE DISENTANGLEMENT OF IMAGE STYLE

Wenya Lu *Zhibin Peng* *Cheng Luo* *Weicheng Xie*†
Jiajun Wen Zhihui Lai Linlin Shen

Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University

ABSTRACT

Image-to-Image synthesis paradigms have been widely used for facial expression synthesis. However, current generators are apt to either produce artifacts for largely posed and non-aligned faces or unduly change the identity information like AdaIN-based generator. In this work, we suggest to use image style feature to surrogate the expression cues in the generator, and propose a multi-task learning paradigm to explore this style information via the supervision learning and feature disentanglement. While the supervision learning can make the encoded style specifically represent the expression cues and enable the generator to produce correct expression, the feature disentanglement of content and style cues enables the generator to better preserve the identity information in expression synthesis. Experimental results show that the proposed algorithm can well reduce the artifacts for the synthesis of posed and non-aligned expressions, and achieves competitive performances in terms of FID, PNSR and classification accuracy, compared with four publicly available GANs. The code and pre-trained models are available at https://github.com/lumanxi236/MTSS.

Index Terms— Facial expression synthesis, multi-task learning, expression style learning, style and content disentanglement

1. INTRODUCTION

Due to the importance of the facial expression for representing human emotion, the task of facial expression synthesis (FES) has attracted increasing attention. Recently, with the progress of GAN-based face editing [1, 2, 3, 4, 5], Image-to-Image (I2I) translation is widely used in FES due to its ability for identity information preservation [3, 4, 6, 7]. While it is difficult to completely separate the expression cues from the identity information, the entanglement of expression and identity is still a major challenge for correctly synthesizing expressions with the identity preservation.

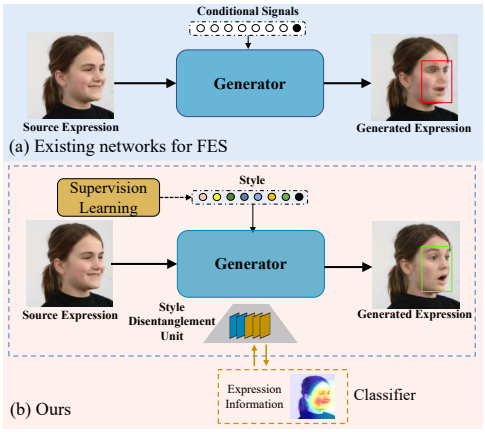


Fig. 1. The existing networks (a) and our proposed network (b) for the task of facial expression synthesis.

Current algorithms resort to the conditional signals, e.g. one-hot label representation [1] for correct synthesis of expression, and the encoder-based feature representation, i.e. skip connection [2, 8, 9] for the generator to better maintain the identity information. However, these conditional signals do not directly reflect the intrinsic feature of each sample, and thus can not accurately represent the cues specific to the complexity of expression variations. Nowadays, AdaIN-based facial attribute editing networks, e.g. L2MGAN [10] and HiSD [11], revealed it reasonable to improve I2I-based attribute editing via the attribute representation with image style. This work motivates us to use the image style in FES to encode the cues specific to the expression characteristics.

However, the style codes learned by AdaIN-based frameworks, e.g. HiSD are too general to accurately represent the specific attribute of expression. In this work, noticing that the learning of facial expression recognition (FER) can make the feature representation be specific to the expression cues, we thus encode more accurate expression style via the supervision learning of FER task.

From another aspect, the generators in the AdaIN-based networks are apt to change the expression-irrelevant informa-

The work was supported by Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grants no. JCYJ20190808165203670, JCYJ20220531101412030.

tion, e.g. the identity cues, during expression editing. To retain the identity cues in FES, most networks devised the regularization losses based on the generated images, e.g. cycle consistency loss [12].

However, due to the intrinsic entanglement of expression and identity cues, image-level constraints may unduly keep the expression features of the source image unchanged while preserving identity cues, which will cause artifacts as shown in the red box in Fig. 1 (a). To this end, we propose a multi-task framework to disentangle the expression and identity cues in the space of feature representation. As shown in Fig. 1 (b), the expression and identity are surrogated with the features of image style and content in the AdaIN-based paradigm. While the disentangled style is mainly used for the expression editing in FES, the remaining identity cues allow the generator to specifically reconstruct the identity details.

To the best of our knowledge, this is the first work to achieve the expression synthesis based on the style-based feature representation of expression cues. Our contributions are summarized as follows

- We propose a supervision-based style encoding (SSE) to encourage the style to be specific to expression cues in the task of FES, under the auxiliary supervision of FER training.
- A style disentanglement unit (SDU) is introduced to separate the expression from identity cues in the feature-level space, which enables the generator to well preserve the identity cues in FES.
- Extensive experiments illustrate the appealing performance of our model in terms of face identity preservation and synthesis quality for posed and non-aligned expressions.

2. THE PROPOSED METHOD

The framework of the proposed algorithm is presented in Fig. 2. Fig. 2(a) presents the supervision-based style encoding (SSE), where the style for representing expression information is encouraged to be more specific with a supervision learning of FER. Fig. 2(b) shows the style disentanglement unit (SDU), where the expression style and identity content are decoupled to not only synthesize realistic expression but also well preserve identity cues.

2.1. Supervision-based Style Encoding (SSE)

To encode the style representation specific to the expression cues, we propose SSE in Fig. 2 (a) by resorting to a reference emotion extractor ($\mathcal{E}$) for encoding the initial image style, an emotion mapper ($\mathcal{M}$) for encoding the expression style provided the expression label, and a classifier ($\mathcal{C}$) for refining the expression style with the supervision training.

First, two random target labels y_{t_1}, y_{t_2}, corresponding to the reference samples of x_{t_1}, x_{t_2}, are selected. $\mathcal{E}$ encodes the

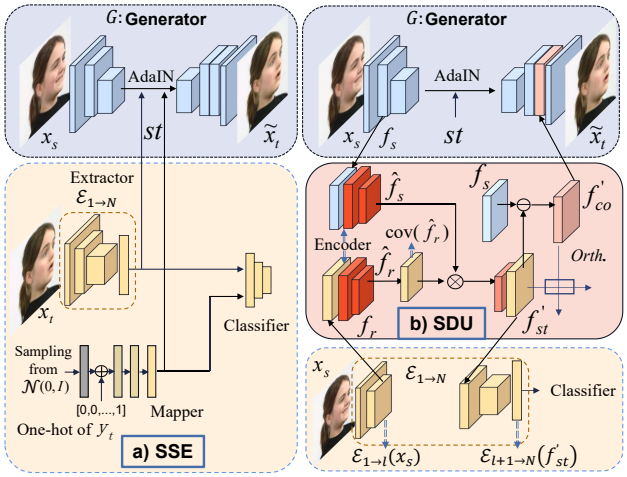


Fig. 2. (a) The supervision-based style encoding (SSE); (b) The style disentanglement unit (SDU). 'Orth.' means orthogonalization.

expression style of x_{t_1}, i.e. $st_1 = \mathcal{E}(x_{t_1})$, and $\mathcal{M}$ generates the style with y_{t_2}, i.e $st_2 = \mathcal{M}(z, y_{t_2})$, where the latent code z obeys the normal distribution, i.e. $\mathcal{N}(0, 1)$. To encourage these styles to be specific to expression cues, expression classifier is used to supervise their feature representation. Specifically, st_1 and st_2 are input into the generator of FES as conditional code, and $\mathcal{C}$ for expression classification as:

$$\mathcal{L}_{SSE} = \mathcal{L}_{CE}(\mathcal{C}(st_1), y_{t_1}) + \mathcal{L}_{CE}(\mathcal{C}(st_2), y_{t_2}) \quad (1)$$

where $\mathcal{C}(st_1)$ outputs a feature representation of st_1 with a MLP, the cross entropy loss $\mathcal{L}_{CE}(\cdot)$ is used for the classifier.

2.2. Style Disentanglement Unit (SDU)

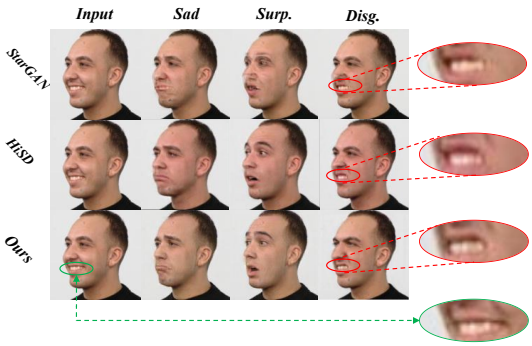
While SSE can learn the styles specific to expression cues, we further decouple the feature to use the content part to supplement the identity of synthesized images. Thus, as shown in Fig. 2 (b), SDU is introduced to further disentangle the expression style and identity content cues, where the style is encoded with the specific channel covariance matrix [13, 14].

First, the features f_s and f_r are obtained as follows

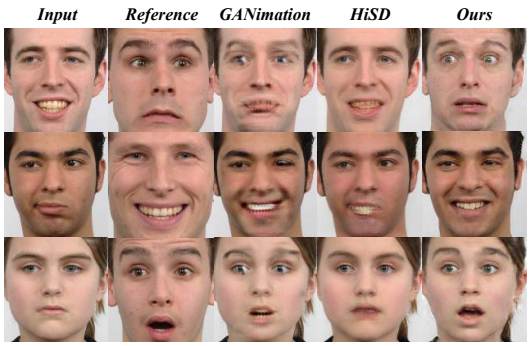
$$\begin{cases} f_s = G_{1\to l}(x_s) \\ f_r = \mathcal{E}_{1\to l}(x_s) \end{cases} \quad (2)$$

where l is the index of an intermediate layer. $G_{1\to l}$ denotes sub-network from the 1-st to l-th layer of the generator in FES. While f_s contains both style (expression) and content (identity) information of x_s, f_r mainly consists of the expression style cues.

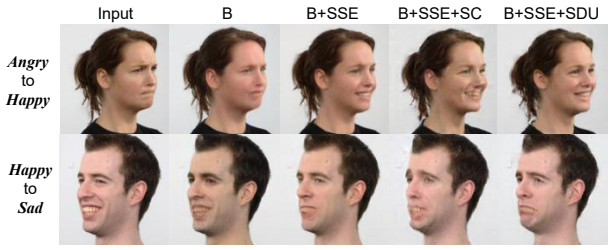
To encode the common expression style information in f_s and f_r, we first perform matrix multiplication on $\hat{f}_s$ and the covariance matrix of flattened $\hat{f}_r$, i.e. $cov(\hat{f}_r)$ to construct their similarity, where $\hat{f}_s$ and $\hat{f}_r$ are the corresponding outputs of f_s and f_r with the weight sharing encoder.



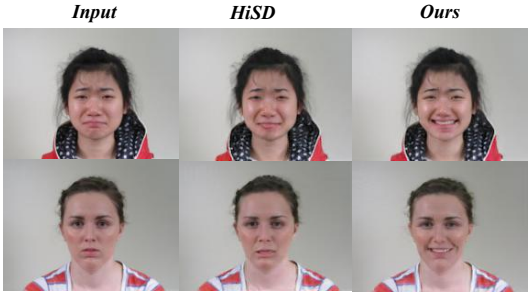
(a) The results for label-guided synthesis on RaFD



(b) The results for reference-guided synthesis on RaFD



(c) Ablation study based on reference-guided results on RaFD



(d) Happy synthesis on non-aligned data of CEFFD

Fig. 3. The visual results, where'B' and 'SC' denote the baseline of HiSD and the skip connection [2].

A decoder is then followed to restore the resulted feature to f'_{st}, i.e. $Dec(\hat{f}_s \otimes cov(\hat{f}_r)$, that has the same dimension with f_s and represents the common style cues contained in f_s and f_r. The identity content features f'_{co} can thus be disentangled from f_s as follows

$$f'_{co} = f_s \ominus f'_{st} = f_s \ominus Dec(\hat{f}_s \otimes cov(\hat{f}_r)) \quad (3)$$

where Dec is the decoder of SDU, $cov(\cdot)$ denotes the covariance matrix, $\otimes$ and $\ominus$ denote the matrix multiplication and the element-wise subtraction. As shown in Fig. 2(b), f'_{co} is then input into the generator of FES for identity detail synthesis.

To make use of SSE in Fig. 2 (a) that can encode expression style information via FER, and further decouple expression and identity features, i.e. f'_{co} and f'_{st} via orthogonalization, we formulate the loss of SDU as follows

$$\mathcal{L}_{SDU} = \mathcal{L}_{CE}(\mathcal{C}(\mathcal{E}_{l+1\to N}(f'_{st})), y_s) + \lambda_o \mathbb{E}||f'_{co} \odot f'_{st}||_1 \quad (4)$$

where $\mathcal{C}$ is the classifier in SSE, N is the number of layers in $\mathcal{E}$, and $\odot$ and $\mathbb{E}$ denote the element-wise product and expectation operator, $||\cdot||_1$ is the L_1 norm, λ_o is a hyperparameter.

2.3. Facial Expression Synthesis (FES)

Adversarial Loss: In addition to the classification loss in Eq. (1), we employ the adversarial loss as:

$$\begin{aligned}\mathcal{L}_{adv} = \mathbb{E}_{x_s, y_s, y_{t_1}, y_{t_2}}[&logD_{y_s}(x_s) + \\ &log(1 - D_{y_{t_1}}(\tilde{x}_{t_1})) + log(1 - D_{y_{t_2}}(\tilde{x}_{t_2}))]\end{aligned} \quad (5)$$

where x_s, y_s, y_{t_1}, y_{t_2}, st_1 and st_2 are shown in Sec. 2.1. $\tilde{x}_{t_1}$ and $\tilde{x}_{t_2}$ are the images synthesized with the generator G as $\tilde{x}_{t_1} = G(x_s, st_1, f'_{co})$ and $\tilde{x}_{t_2} = G(x_s, st_2, f'_{co})$.

Consistency Loss: In addition to the cycle consistency loss [12], i.e. $\mathcal{L}_{cyc}$, we introduce a consistency loss to make the generator insensitive to the inputs, i.e. narrowing the generation gap when image x_s or $\tilde{x}_{t_1}$ is input into G. The consistency losses are formulated as:

$$\mathcal{L}_{rec} = \mathcal{L}_{cyc} + \mathbb{E}_{x_s, \tilde{x}_{t_1}, \tilde{x}_{t_2}}||G(\tilde{x}_{t_1}, st_2, f'_{co}) - \tilde{x}_{t_2}||_1 \quad (6)$$

Style Reconstruction Loss: In order to limit the generated image to contain the target expression style, we resort to a style reconstruction loss, i.e. $\mathcal{L}_{st}$ that was used in [5, 15, 16].

Overall Loss: The overall loss is formulated as:

$$\mathcal{L}_{all} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{st}\mathcal{L}_{st} \quad (7)$$

where $\lambda_{adv}, \lambda_{rec}, \lambda_{st}$ are hyperparameters.

3. EXPERIMENT

Dataset. For the evaluation, Radboud Faces Dataset (RaFD) [17] and Compound Facial Expressions of Emotions Dataset (CFEED) [18] are employed.

Baseline networks. The popular networks of StarGAN [1], GANimation [3], HiSD [11] and EF-GAN[4] are used as the baselines.

Hyperparameter setup. In Eq. (2) of SDU, $l = 3$. All the models are trained for 300k iterations. $\lambda_{adv}, \lambda_{rec}, \lambda_{st}$ in Eq. (7) are set as 1. λ_o in Eq. (4) is set as 0.001.

3.1. Visualization and Quantitative Analysis

Target label-guided synthesis. To test the performance of the proposed expression synthesis, we show the synthesized results for posed faces in Fig. 3 (a). Fig. 3 (a) shows that StarGAN produces apparent artifacts in the posed image. HiSD can synthesize target expression correctly, while the skin color and tooth texture have unduly changed. By contrast, our model can not only synthesize target expressions correctly but also well preserve these identity details.

Reference-guided synthesis. To test the performance under the reference image guidance, Fig. 3 (b) presents the results of three algorithms, where the samples generated by GANimation appear artifacts, and HiSD generates failure cases, i.e. the skin color is interfered by the reference image. By contrast, our method can well synthesize the expressions specific to the reference image with the preservation of face color.

Quantitative Results of FES. To quantitatively evaluate the performances of the proposed algorithm, Fréchet Inception Distance (FID) [19] is used to test the realism of the synthesized image, and peak-signal-to-noise ratio (PSNR) [20] is used to test the preservation of identity information. Classification accuracy (CA) is evaluated on the generated expressions with Resnet18 [21], which is pre-trained on the training set of RaFD. The results on RaFD are shown in Tab. 1.

Tab. 1 shows that our method can generate samples with higher quality in terms of FID and PSNR for both label-guided and reference-guided synthesis. Especially, in the non-alignment images, i.e. the 2nd column, our algorithm achieves large improvements of 5.35 in terms of FID for 'R' synthesis. In terms of CA, our method ranks the 2nd among four algorithms. Resorting to an auxiliary expression classifier [22] during training, StarGAN generates easily distinguishable expressions and achieves the best CA, while the identity preservation is not satisfactory, as shown in Fig. 3 (a). For reference-guided synthesis, our method achieves much better performance than that with HiSD.

In addition to the performance on challenging posed faces, we also study the performance of our algorithm specially on frontal faces in Tab. 1. The last two rows in Tab. 1 show that our model outperforms EF-GAN [4] in terms of all the three metrics, based on the same protocol of EF-GAN, i.e. only the frontal faces are used.

Ablation study. In order to study the effects of SSE and SDU, visual result and quantitative result of an ablation study based on the reference-guided synthesis are presented in Fig. 3 (c) and Tab. 2.

By comparing the 2nd and 3rd columns in the 1st row of Fig. 3 (c), it shows that SSE enables the baseline of HiSD to generate the target expression, while the identity cues are not well preserved. By comparing the 3rd or 4th column with the 5th column, it shows that the additional SDU enables the generator to better preserve the identity details. Tab. 2 shows that SDU can further improve the quality of the synthesized image, identity information preservation, and the accuracy of target expression editing.

Table 1. The results in terms of FID, PSNR and CA. 'R' and 'T' refer to reference- and label-guided synthesis, 'Cp' means that the data is processed with cropping, 'F' and 'P' denote the frontal and posed faces, respectively. GANima is the abbreviation of GANimation.

Model	Data	FID(↓)	FID(Cp)	PSNR(↑)	CA(↑)
StarGAN (T)	F&P	45.31	-	13.44	**93.45**
GANima (R)	F&P	-	23.99	-	77.01
HiSD (R)	F&P	49.30	27.91	13.08	22.02
HiSD (T)	F&P	44.05	22.25	13.03	92.44
Ours (R)	F&P	43.95	**22.19**	13.63	92.26
Ours (T)	F&P	**42.52**	22.60	**13.80**	93.38
EFGAN (R)	F	42.36	-	23.07	89.38
Ours(R)	F	**36.19**	-	**25.75**	**92.88**

Table 2. The ablation results in terms of FID, PSNR and CA.

Model	FID(↓)	PSNR(↑)	CA(↑)
Baseline (B)	49.30	13.08	22.02
B+SSE	47.26	12.83	90.08
B+SSE+SC	45.24	12.99	89.78
B+SSE+SDU	**43.95**	**13.63**	**92.26**

Synthesis on non-aligned faces To evaluate the performance of the proposed algorithm on more challenging task, i.e. non-aligned expression synthesis, i.e. the image used for testing is not cropped. We present the results of 'happy' expression synthesis on CFEED dataset in Fig. 3 (d).

Fig. 3 (d) shows that HiSD fails to synthesize the target expressions correctly for the samples in the 1st and 2nd columns. By contrast, our algorithm correctly synthesizes the 'happy' expression and well preserves the identity cues.

4. CONCLUSION

To reduce the artifacts and preserve the identity cues in facial expression synthesis (FES), this work newly employs the image content and style to surrogate the identity and expression cues, respectively. While the style is supervised learned with expression recognition to make it more specific to the expression cues, the image content after disentanglement enables the generator to better preserve the identity in FES. Experimental results show that our algorithm largely outperforms four state-of-the-art GANs in terms of synthesis quality and identity preservation. We will develop smooth distribution of style code to improve our model for the application of facial expression animation.

5. REFERENCES

[1] 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, pp. 8789–8797.

[2] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE TIP*, vol. 28, no. 11, pp. 5464–5478, 2019.

[3] 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, pp. 818–833.

[4] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *CVPR*, 2020, pp. 5021–5030.

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *CVPR*, 2020, pp. 8188–8197.

[6] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *CVPR*, 2018, pp. 7083–7092.

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

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[10] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang, "L2m-gan: Learning to manipulate latent space semantics for facial attribute editing," in *CVPR*, 2021, pp. 2951–2960.

[11] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji, "Image-to-image translation via hierarchical style disentanglement," in *CVPR*, 2021, pp. 8639–8648.

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[13] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang, "Universal style transfer via feature transforms," *NeurIPS*, vol. 30, 2017.

[14] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai, "CCPL: Contrastive coherence preserving loss for versatile style transfer," in *ECCV*, 2022.

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018, pp. 35–51.

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.

[17] 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.

[18] Shichuan Du, Yong Tao, and Aleix M Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017.

[20] Alain Hore and Djemel Ziou, "Image quality metrics: Psnr vs. ssim," in *ICPR*, 2010, pp. 2366–2369.

[21] S. Ren Kaiming He, X. Zhang and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[22] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier gans," in *ICML*, 2017, pp. 2642–2651.