DisFlowEm : One-Shot Emotional Talking Head Generation using Disentangled Pose and Expression Flow : Technical Appendix

1. Experimental Details

1.1. Network Architecture Details

1.1.1 Expression Generation Network:

The expression generation network G_{exp} consists of an Audio Encoder E_a, Emotion Encoder E_e, Mouth Graph Encoder E_{mg}, Face Graph Encoder E_{fg}, and Face Graph Decoder D_{fg}. *Audio Encoder* E_a uses an input emotion-invariant DeepSpeech [3] features of size $R^{6\times 29}$ corresponding to each video frame. E_a consists of 3 LSTM layers which with hidden size 256. The output of E_a consists of an audio feature vector of length 128. *Emotion Encoder*, E_e consisting of a single convolutional layer, encodes an input emotion (one-hot vector consisting of six emotion labels and two different intensities) to an emotion feature vector of length 128. *Face Graph Encoder,* E_{fg} encodes the input canonical frontal face landmark graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ using spectral graph convolution [5] to a face graph feature vector of length 128 using hierarchical graph convolution [9]. A *Face Graph Decoder* D_{fg} reconstructs the output landmark graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}, \mathcal{A})$ from the concatenation of the feature vectors f_a, f_l, f_e by performing graph upsampling. In contrast to [9] our proposed method uses a *Mouth Graph Encoder* E_{mg} which performs graph convolution of mouth landmark graphs $\mathcal{G}^{\Updownarrow} = (\mathcal{V}^m, \mathcal{E}^m, \mathcal{A}^m)$ consisting of mouth landmark vertices for improving lip sync accuracy with the help of a landmark lip sync loss L_{sync} (Eq. 3 in main paper).

1.1.2 Pose Generation Network:

The proposed Pose Generation Network as shown in Fig. 3 of the main paper, consists of an Audio Encoder E_a, Emotion Encoder E_e, and Pose Encoder E_p. The Audio Encoder E_a uses an LSTM network of 3 layers (hidden size 256) to encodes DeepSpeech features to an audio feature vector of length 128. The Emotion Encoder E_e encodes the input one-hot emotion vector to a fixed feature vector of length 128. The Pose encoder also encodes the reference pose and the input ground truth pose (during training) independantly. These input features are concatenated and passed to a bidirectional LSTM network consisting of 2 hidden layers of size 256. The decoder is another bidirectional LSTM of (2 hidden layers of size 256) which decodes the noise variable Z (sampled from N(μ,σ) during training) conditioned on the input emotion, audio and pose reference features. The output feature produced by the decoder LSTM (size 256) is passed through a linear FC layer to get a predicted pose displacement vector of dimension 6.

1.1.3 Image Generation Network

Input identity image I_{id} of size 256x256x3 is first downsampled to 64x64x3. This image is warped using TPS transformations for expression and pose respectively to generation I_e and I_p respectively. I_e is concatenated along the channel dimension with difference of gaussian heatmaps (h_m) computed between expression landmarks and neutral landmarks $h_m(L_{exp}) - h_m(L_{neu})$. This concatenated feature is used to generate dense optical flow and occlusion map for expression branch, using an hourglass network as shown in Fig. 4 of the main paper. Similarly I_p is concatenated along the channel dimension with $h_m(L_{pose}) - h_m(L_{neu})$, and sent to the pose branch of motion generation network to compute pose flow and occlusion map. Each hourglass network consists of 5 downsampling blocks followed by 5 upsampling blocks. The output of each hourglass network is followed by a convolutional layer that generates flow and occlusion maps of size 64x64x2 respectively. In the Image reconstruction stage, the identity image 256x256x3 is encoded by an identity encoder consisting of 4 convolution+downsampling layers. The decoder upsamples the bottleneck layer feature to an inpainting image of resolution 256x256x3. To preserve the identity information of the source image in the inpainted image, the downsampled feature maps after each layer are warped using the occlusion and flow map of the respective branch and then concatenated with the decoder feature map at the previous resolution. The expression branch additionally uses an emotion input concatenated with the lowest resolution feature map before being sent to the decoder.

1.2. Single-stage optical flow-based Texture Generation (for Ablation)

Ablation configuration (1) Our method w/o disentangled learning: In order to emphasize the importance of two branch generation, we also train a model with a *single branch* for computing optical flow based on expression and pose landmarks. The Motion Generation Network learns the emotion-conditioned facial motion with the help of a dense optical flow map and occlusion map [4, 15]. The optical flow map indicates the transformation of the source image I_{id} to generate the final output image. The occlusion map indicates the regions that are occluded in the source image which need to be inpainted in the generated image.

$g : (I_{id}, tps_e, tps_p, h_m(L_{exp}), h_m(L_{pose}), e) \rightarrow (flow, occ)$

The flow map is generated as follows, $flow = M * tps_e + (1 - M) * tps_p$, where M is a pixel contribution map that is used to combine the TPS transformation function due to expression and pose deformation parameters. The optical flow map is used to warp the feature maps of I_{id} encoded by the inpainting network. The final generated image is obtained as follows :

$I_{exp+pose} = \tau(I_{id}, flow) * occ + I_{inpaint} * (1 - occ)$

where $I_{inpaint}$ is the texture map containing inpainted texture of the face.

The single-flow based texture generation network is finetuned on MEAD dataset to learn emotional talking face. However due to the low pose variety of MEAD, the network is unable to retain the head pose variety learnt during pretraining on large scale emotion-agnostic datasets, due to a single combined optical flow and occlusion map. While the face moves the hair remains static, as visible in the ablation section of the supplementary video, and Fig. 8 of main paper.

1.3. Data-preprocessing:

The ground truth videos of training datasets HDTF [15], MEAD [10] and RAVDESS [6] recorded at 30 fps are cropped to 256x256 size frames. The identity image I_{id} is in fixed frontal (or near frontal) pose which is aligned to frontal face during evaluation. The ground-truth landmarks are extracted using 3DDFA [2] and [13] following [8]. The ground truth poses are computed by computing a rigid transformation from the neutral pose identity image landmark L_{neu}. The ground truth landmarks are aligned via procrustes alignment [1] to canonical face landmarks in a fixed frontal pose in order to train G_{exp}, following [9]. DeepSpeech [3] features are computed from the input audio corresponding to each video frame. During fine-tuning of expressions branch of G_{image} on MEAD and RAVDESS, due to the poor variability of these datasets we perform background augmentation and random colour and brightness manipulation to allow variability during training. This helps in improved generalization to arbitrary faces and backgrounds at test time.

1.4. Implementation Details and Training Strategy

The reference pose p_0 used in G_{pose} is extracted from the input image in frontal or near frontal pose. The CVAE-LSTM network generates sequences of 25 frames. After each sequence the reference pose is initialized with the last generated pose in the previous sequence. For each pose in the sequence G_{pose} generates pose deviation $\delta\hat{p}$ from the reference pose. The pose is represented by a six-dimensional pose vector consists of euler rotation and translation coefficients. The final predicted pose is used to transform the neutral landmark L_{neu} via a rigid alignment (using nose landmarks) to *pose landmarks* L_{pose}. G_{pose} is trained using GT pose supervision from dataset RAVDESS [6] since it has slightly higher variations in head pose than MEAD.

The Image Generation Network G_{image}, as shown in Fig. 4 in the main paper is trained in two stages. In the pre-training stage, the entire network G_{image} is trained on the large-scale [15] dataset which contains mostly neutral and smiling faces, but does not contain emotion labels. The emotion label input is randomly assigned during the pre-training to reduce biasness towards any given emotion. During the expression finetuning stage, the network layers of the pose branches of the G_{exp} and the identity encoder of the expression inpainting branch are frozen, while remaining layers are finetuned on the emotional training dataset.

G_{image} is trained using Adam Optimizer with an initial learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and weight decay of 0.0001, and batch size of 24 on a aingle 48 GB NVIDIA A100 GPU. G_{exp} and G_{pose} are trained using Adam Optimizer with an initial learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.0004 and batch size 100 respectively. G_{image} is pre-trained for 25 epochs for 8 days on a single NVIDIA A100 GPU on 210 subjects from large scale dataset HDTF [15]. Then the expression flow-guided branches of G_{image} are finetuned for 5 epochs (around 6 hours) on emotional training data consisting of total 60 subjects from MEAD [10] and RAVDESS [6]. Training G_{exp} takes around a day with batch size 256. Training G_{pose} takes around 4 hours with batch size 100.

2. Justification for Emotion-controllable Head movements

Table 1 demonstrates the standard deviation of head movements on ground truth videos from RAVDESS dataset (computed by the displacement of the nose tip landmark over the entire video). It can be seen that the degree of head movements is lowest for neutral emotion, and the degree of head movements is higher for happy and neutral. This

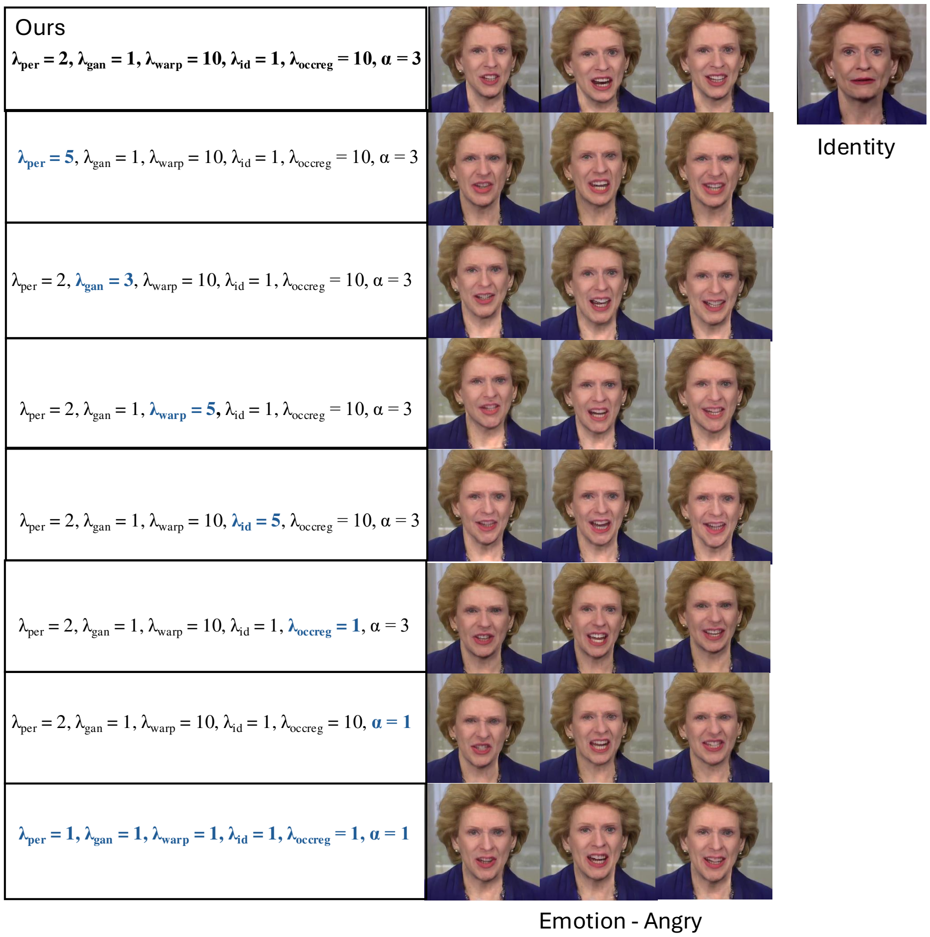
Ours
$\lambda_{per} = 2, \lambda_{gan} = 1, \lambda_{warp} = 10, \lambda_{id} = 1, \lambda_{occreg} = 10, \alpha = 3$
$\lambda_{per} = 5, \lambda_{gan} = 1, \lambda_{warp} = 10, \lambda_{id} = 1, \lambda_{occreg} = 10, \alpha = 3$
$\lambda_{per} = 2, \lambda_{gan} = 3, \lambda_{warp} = 10, \lambda_{id} = 1, \lambda_{occreg} = 10, \alpha = 3$
$\lambda_{per} = 2, \lambda_{gan} = 1, \lambda_{warp} = 5, \lambda_{id} = 1, \lambda_{occreg} = 10, \alpha = 3$
$\lambda_{per} = 2, \lambda_{gan} = 1, \lambda_{warp} = 10, \lambda_{id} = 5, \lambda_{occreg} = 10, \alpha = 3$
$\lambda_{per} = 2, \lambda_{gan} = 1, \lambda_{warp} = 10, \lambda_{id} = 1, \lambda_{occreg} = 1, \alpha = 3$
$\lambda_{per} = 2, \lambda_{gan} = 1, \lambda_{warp} = 10, \lambda_{id} = 1, \lambda_{occreg} = 10, \alpha = 1$
$\lambda_{per} = 1, \lambda_{gan} = 1, \lambda_{warp} = 1, \lambda_{id} = 1, \lambda_{occreg} = 1, \alpha = 1$
Identity
Emotion - Angry

Figure 1. **Ablation study of loss parameters of Texture Generation network. We have presented qualitative and quantitative ablation by removing each of the loss term in the texture generation in Fig. 8 and Table 3 in the original paper. The loss importance parameters in the paper were heuristically chosen for best results (identity preservation, texture quality) based on random search. Our method is not extremely sensitive to the choice of loss importance weights, as indicated by the results.**

motivates us to learn emotion-controllable head movements which is not explored in prior works.

Emotion	SD
neutral	6.75
angry	38.35
disgusted	24.25
fear	28.23
happy	33.33
sad	18.41
surprised	16.89

Table 1. Standard deviation of head displacements on RAVDESS dataset.

3. Additional Results

Results on neutral emotion: To demonstrate that the performance of our model pre-trained on neutral emotion is comparable with state-of-the-art talking head methods in neutral emotion, we present quantitative results on HDTF test data in Table 3 and qualitative results in Fig. 3. Our method achieves the best value for texture quality metric FID, and also outperforms existing methods in identity preservation, indicated by the highest CSIM value. Although SadTalker achieves the highest lip sync accuracy, our M/F-LMD is lower indicating good lip sync in terms of facial landmarks.

4. Limitations and Future Scope

One limitation of our method is that retaining the generalization ability after expression finetuning is challenging owing to the limited variety of the emotional training data. Hence early stopping is crucial to prevent overfitting on the smaller training set of emotional data. The disentangled pose and expression flow computation does not

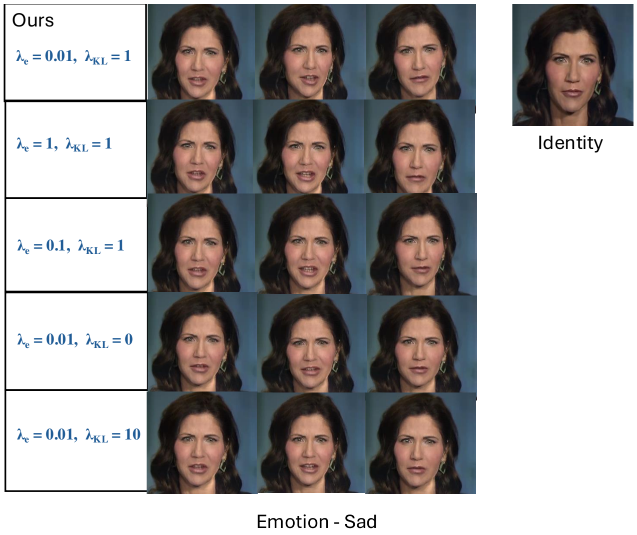
Ours
$\lambda_e = 0.01,\ \lambda_{KL} = 1$
$\lambda_e = 1,\ \lambda_{KL} = 1$
$\lambda_e = 0.1,\ \lambda_{KL} = 1$
$\lambda_e = 0.01,\ \lambda_{KL} = 0$
$\lambda_e = 0.01,\ \lambda_{KL} = 10$
Identity
Emotion - Sad

Figure 2. **Ablation study of loss parameters of Pose Generation network. The chosen configuration results in higher pose diversity.**

Method	SSIM ↑	FID ↓	M/F-LMD ↓	CSIM ↑	Sync$_{conf}$ ↑
MakeItTalk [17]	0.593	28.243	4.45/5.08	0.838	2.563
Wav2Lip [7]	0.618	21.725	3.63/4.54	0.849	5.227
PC-AVS [16]	0.422	69.127	3.93/10.51	0.683	2.701
Audio2Head [11]	0.60	24.392	2.48/8.34	0.823	3.90
AVCT [12]	0.755	22.432	3.61/2.73	0.811	3.147
FGTF [15]	**0.840**	-	**0.39**/-	-	5.166
SadTalker [14]	0.532	22.057	2.39/2.01	0.843	**7.290**
Ours	0.798	**14.70**	0.82/**0.73**	**0.856**	5.287

Table 2. Quantitative Comparison with one-shot **neutral emotion** methods on neutral emotion dataset HDTF. The best value of a metric is marked in **Bold** and the second best is marked in **Blue**. Note: Wav2Lip and PC-AVS are evaluated using a single identity reference image. Since the full code/pre-trained models for FGTF [15] are not available, we report the available metrics from their paper.

take into consideration any temporal constraints, which is another limitation of our method. Future directions might be in adding temporal constraints in the disentangled optical flow generation for smoother temporal transitions in the synthesized animation.

References

[1] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, 2020. 2

[2] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. `https://github.com/cleardusk/3DDFA`, 2018. 2

[3] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 1, 2

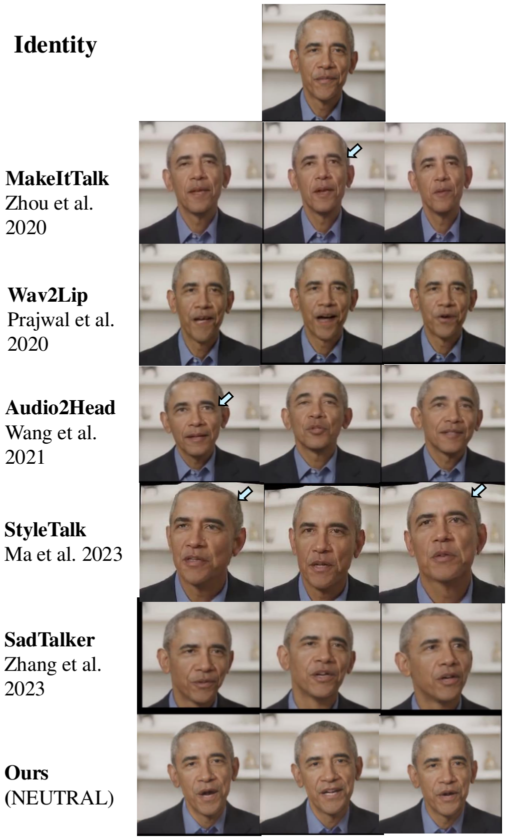

Figure 3. **Qualitative Comparison with one-shot talking head generation methods on neutral emotion dataset HDTF.** Our method is able to generate identity-preserving facial texture with different head movements for different emotions. MakeItTalk, Wav2Lip, Audio2Head, SadTalker generate talking heads only in neutral emotion. StyleTalk derives the speaking style from a smiling person's talking video. MakeItTalk, Audio2Head are not able to accurately capture the identity.

[4] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021. 2

[5] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[6] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 2

[7] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 4

[8] Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. Identity-preserving realistic talking face generation. In

2020 International Joint Conference on Neural Networks (IJCNN), pages 1–10. IEEE, 2020. 2

[9] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *International Joint Conference on Artificial Intelligence*, 2022. 1, 2

[10] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 2

[11] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. 2021. 4

[12] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022. 4

[13] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018. 2

[14] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 4

[15] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 2, 4

[16] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 4

[17] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 4