ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747284

META TALK: LEARNING TO DATA-EFFICIENTLY GENERATE AUDIO-DRIVEN LIP-SYNCHRONIZED TALKING FACE WITH HIGH DEFINITION

Yuhan Zhang[1*], *Weihua He*[1*], *Minglei Li*[2], *Kun Tian*[1], *Ziyang Zhang*[1#], *Jie Cheng*[1], *Yaoyuan Wang*[1#], *Jianxing Liao*[1]

[1] Advanced Computing and Storage Lab, Huawei Technologies Co. Ltd., China
[2] Language & Speech Innovation Lab, Huawei Technologies Co. Ltd., China

ABSTRACT

Audio-driven talking face, driving talking face by audio, has received considerable attention in multi-modal learning due to its widespread use in virtual reality. However, long-time recording of target high-quality video is needed by most existing audio-driven talking face studies, which significantly increases customization costs. This paper proposes a novel data-efficient audio-driven talking face generation method, which uses just a short target video to produce both lip-synchronized and high-definition face video driven by arbitrary audio in the wild. Current methods suffer from many problems, such as low definition, asynchronization of lip movement and voice, and intense demands for videos for training. In this work, the original target character's face images are decomposed into 3D face model parameters including expression, geometry, illumination, etc. Then, low-definition pseudo video generated by an adapted target face video bridges the powerful pre-trained audio-driven model to our audio-to-expression transformation network and help to transfer the ability of audio-identity disentanglement. The expression is replaced via an audio and then combined with other face parameters to render a synthetic face. Finally, a neural rendering network translates the synthetic face into talking face without loss of definition. Experimental results show that the proposed method has the best performance in high-definition image quality, and comparable performance in lip synchronization compared with the existing state-of-the-art methods.

Index Terms— Talking face generation, Lip sync, High definition, Audio driven animation

1. INTRODUCTION

Talking face generation has attracted intensive attention in the field of multi-modal human-computer interaction. Its powerful communication mode makes it widely used in virtual education, videoconferencing, virtual anchor, game entertainment, film and television animation production and other fields. Accurate lip movements during lip-syncing and realistic video portraits are key to better user experience feedback [1, 2, 3, 4, 5], natural head poses and eye blinks can also enhance user experience in these application scenarios [6, 7, 8]. However, The existing talking face generation methods [1, 2, 3, 4, 5] can not simultaneously achieve high definition and lip synchronization. How to realize them and develop audio-driven talking face generation has aroused the interest of many researchers.

The traditional image-based approach [1] requires a large number of target character's videos because it retrieves the best matching lip image from the database and splices it back to the original background image. The method based on 3D morphable model (3DMM) [2] reduces the burden of target video acquisition, but lip shape is not synchronized well with an arbitrary new piece of audio in the generated video as its audio-driven performance strongly depends on the audio identity. The approach based on GAN [4] uses a pre-trained discriminator to accurately detect lip-sync errors and force the generator to accurately morph the lip movements in sync with a new audio in the wild instead of the target's audio. Although it produces a decent lip-syncing video of the talking face and achieves disentanglement of audio identity and model and target identity, the definition of the lip area is always poor for visual experience and cannot meet application requirements.

In this paper, we propose a novel talking face generation framework and strive to transfer the powerful audio-driven lip-syncing abilities from a pre-trained model [4] to ours using only short training target video. The key contribution of this paper are summarized as follows:

- In the proposed method, the low-definition pseudo video predicted by Wav2Lip[4] with the target video and LRS2 audio [9] is introduced to enhance the audio-driven identity-disentangled ability of talking face generation.
- We train a modified audio-to-expression (A2E) network to guarantee the accurate lip motion driven by arbitrary audio, which makes our method possess an powerful audio-driven performance comparable to Wav2Lip[4].
- A modified crop module is introduced for automatically adapting the size of the 3DMM synthetic face to the original face area, then enabled our framework to meet the requirements of 4K-definition photo-realistic talking face video.

2. OUR METHOD

Fig.1 summarizes the pipeline of the proposed method. During the training phase, we first crop the original target video

*Equal contribution.
#Corresponding Author: zhangziyang11@huawei.com, wangyaoyuan1@huawei.com

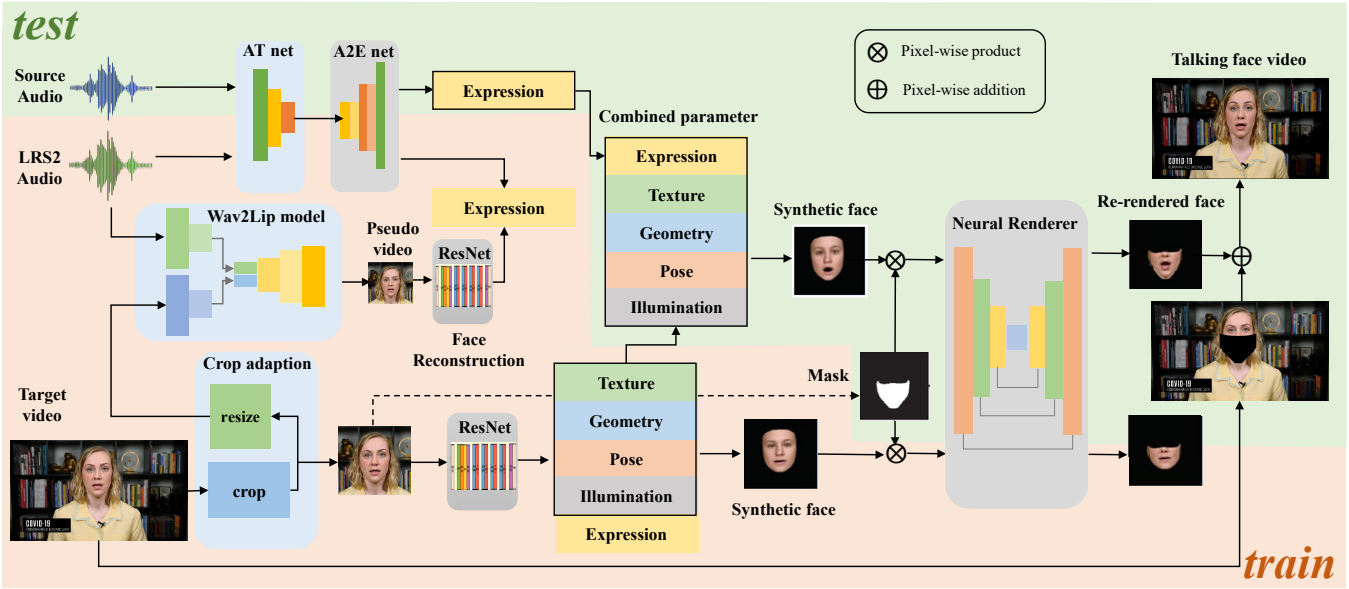

Fig. 1. The framework of our method including training and testing processes. Modules marked in gray require training, but modules marked in blue do not.

into the target face video, which is then resized to be low-resolution to generate a low-definition talking face video with LRS2 [9] audio using the pre-trained model Wav2Lip [4]. The generated video-audio pair is the pseudo label which possesses abundant phonemes and corresponding talking face video with excellent lip synchronization. 3D face reconstruction is performed on both the pseudo video and the target video, and the facial 3D morphable model (3DMM) parameters including expression, geometry, texture, pose, illumination coefficients are extracted from each frame of them. To obtain a powerful model mapping audio to expression parameters, a new audio-to-expression transformation network is trained with audio-expression pairs of the pseudo video. Then, the 3DMM parameters are used to re-render the synthetic facial images in the target video. Finally, we train a neural rendering network with the lower half of synthetic and real target faces to generate a high-definition photo-realistic talking face video.

During the testing phase, arbitrary audio can be input and fed into the trained audio-to-expression transformation network to predict audio-driven expression parameters. Then, the predicted expression parameters replace the original ones of 3DMM parameters obtained by 3D reconstruction. We re-render the face to audio-driven synthetic face using the combined 3DMM parameters. Then the lower half of the generated synthetic face is translated to a realistic lower half face. Finally, the generated photo-realistic lower half face is sewn into the background of original target video to generate a high-definition lip-synchronized video.

2.1. Crop adaptation

Given the target video, we first detect its face and crop it into a target face video with the image size of 512x512 or 768x768 or 1024x1024. The image size is adapted to the face resolution of the original video. This step, followed by subsequent 3D reconstruction and neural face renderer generation, can generate a face with the same resolution as the target face, which facilitates HD face translation. Then, the cropped face video is resized to a low-resolution video with size of 256x256. We feed this video together with LRS2 audio [9] into the pre-trained talking face generation model [4] to obtain the pseudo video. Predicted by Wav2Lip, the pseudo video has low definition and good lip synchronization. In the proposed method, we perform 3D reconstruction on the pseudo video to obtain expression parameters and train the A2E network, and transfer the powerful lip synchronization performance of wav2lip to our A2E network. So the pseudo video is an important link between our approach and Wav2Lip. The resize operation reduces the size of the pseudo video but can speed up the 3D reconstruction significantly. Our method focuses on driving the expression parameters obtained from 3D face reconstruction via audio, which does not depend on the frame size of the input video, so we can use low-definition pseudo-video to extract the expression parameters.

2.2. 3D Face Reconstruction

Fig. 2. List of target characters' videos for training in our experiments.

In order to generate audio-driven talking face, 3D face model should be reconstructed first for the target face video and pseudo video. And, expression, geometry, texture, illumination, pose coefficients are extracted from each frame of them, which can fully depict the facial performance of the

video and facilitate to render back to synthetic face image. The 3D face reconstruction could be performed on videos with low frame resolution and definition due to the priori embbeding of 3DMM, leading to 3D face parameter estimation. We utilize the deep-learning-based face reconstruction model [10] where the input face I is represented as 3DMM, a parametric face model. Then a deep CNN is trained to estimate the 3DMM parameters Φ_I from the input face I. The face geometry shape s can be reconstructed as $s = \overline{s} + \alpha^T \Xi_{exp} + \beta^T \Xi_{geo}$, and the facial reflectance r is formulated as $r = \overline{r} + \gamma^{\mathbf{T}} \Xi_{ref}$ to depict facial texture, where $\overline{s}$ and $\overline{r}$ are represented as the average facial shape and reflectance, $\alpha \in \mathbb{R}^{64}, \beta \in \mathbb{R}^{80}, \gamma \in \mathbb{R}^{80}$. $\Xi_{exp}, \Xi_{geo}, \xi_{ref}$ denote the matrix of expression, geometry and reflectance basis, respectively. They are all calculated from facial scan data by principal component analysis (PCA). The Basel face model of 2009 [11] is adopted for $\overline{s}, \overline{r}, \Xi_{geo}$ and Ξ_{ref}. And, CNN-based coarse-to-fine learning approach [12] is used to obtain Ξ_{exp} based on Face-Warehouse [13]. Moreover, we approximate environment lighting using spherical harmonics (SH) [14] with the assumption of Lambertian surface and distant scene illumination to model the illumination. The radioactivity of the vertex μ_i with normal k_i and texture z_i can be computed as $C(k_i, z_i, \delta) = z_i \sum_{b=1}^{B^2} \delta_b \psi_b(k_i)$, where $\psi_b : \mathbb{R}^3 \to \mathbb{R}$ are SH basis functions, δ_b are SH coefficients and $B = 3$ bands of SH. These results in SH coefficient $\delta \in \mathbb{R}^{27}$. Pinhole camera is used to model the virtual camera for perspective projection from 3D to 2D. The rigid head pose is represented by rotation $R \in SO(3)$ and translation $T \in \mathbb{R}^3$. The complete 3DMM is represented by $\Phi_I = \{\alpha^T, \beta^T, \gamma^T, \delta^T, R^T, T^T\}^T \in \mathbb{R}^{257}$. We adopt the same finetuning step as AudioDVP [2] to finetune the 3D reconstruction network for both the pseudo video and the target video.

2.3. Audio to Facial Expression Mapping

The performance of lip sync depends heavily on the construction of the mapping from audio signal to facial expression, as it is known that the movements of the lip and the lower face are strongly correlated with the audio signal. A rich training corpus of audio-expression pairs is an important prerequisite for good arbitrary-audio-driven performance of the mapping model. As LRS2 dataset [9] is a large-scale spoken sentences corpus from BBC television and Wav2Lip [4] is a powerful lip-sync generation model trained on LRS2, the pseudo video is lip synced well according to our extensive literature and user research experiments. It is full of abundant phonemes and lip-synced video paired with them. Therefore, We use LRS2 audio [9] and 3D facial expression estimated from the pseudo video predicted by the pre-trained generation model Wav2Lip [4] to build an effective audio-to-expression transformation network.

First, the MFCC feature of the input audio is extracted and fed into AT-net[3] to obtain the 256-D high-level feature f. Then, an A2E network is established to map this feature to the paired 3D facial expression parameters. Due to the large and rich training data of audio-expression pairs, we deepen the A2E network in AudioDVP [2] in order to enhance the modeling and fitting capabilities of the network (see Table 1).

Table 1. Architecture of our A2E network.

Layer	Conv1D	Conv1D	Conv1D	Conv1D	Conv1D	FC
Kernel	3	3	3	3	3	-
Stride	1	1	1	1	1	-
Outputs	5x254	5x252	3x250	3x248	3x246	64

To train the A2E network H, the mean squared error (MSE) loss L_{A2E} is formulated as follows:

$$L_{A2E} = MSE(H(f) - \alpha) \tag{1}$$

where α is the expression coefficient estimated from the pseudo video.

2.4. Neural Face Rendering Network

After the 3D reconstruction, the synthetic face image can be rendered using the estimated 3DMM coefficients. In order to obtain high-quality natural-looking face image, the synthesized face image is translated into high-definition, photorealistic face image.

The masking strategy of AudioDVP [2] is adopted to obtain the lower half of the synthetic face $\widetilde{I}$ and the target face I, which form the training corpus for neural face rendering translation. The synthesized rendered face is translated into a photorealistic rendering of the target face by the neural rendering network. The neural rendering network consists of a U-Net-based generator G and a discriminator D. We adapt the resolution of the target face by modifying the input size of the U-net-based face generator. The generator consists of a face encoder and a face decoder. The rendering face encoder is a stack of downsampled convolutional layers, which encodes the lower half of the synthetic face and obtain an advanced feature representation of it. The feature is then upsampled by the transposed-convolution-stacked decoder to synthesize high-quality outputs. The discriminator uses PatchGAN [15], and the training generator minimizes the L_1 reconstruction loss of the generated rendered face $G(\widetilde{I})$ and the ground-truth face I:

$$L_{rec}(D) = ||I - G(\widetilde{I})||_1, \tag{2}$$

The input to the discriminator D is the ground-truth face I and the rendered picture $G(\widetilde{I})$. The loss of the GAN is

$$L_{adv}(G, D) = logD(I) + log(1 - D(G(\widetilde{I})), \tag{3}$$

Therefore, the loss function of the entire network is

$$L(G, D) = L_{rec}(D) + \lambda L_{adv}(G, D). \tag{4}$$

3. EXPERIMENTS

We tested our method on the videos of seven characters (see Fig.2) collected from the previous work [2]. Only $3min$ of them are used to train the model. We first aligned all the speaking faces by detecting their landmarks, and then cropped the video to a 512x512 or 768x768 frame size centered around the lower half of the face. For audio data, similar to [3], we used a window of $10ms$ in size to extract the MFCC. Then,

Table 2. Quantitative evaluation on the test sets of videos. For LSE-D and FID the lower the better, and the higher the better for LSE-C and SSIM.

	Methods	ATVG	Wav2Lip	AudioDVP	MakeIttalk	Ours
A	LSE-D↓	9.114	7.756	10.195	9.977	8.636
	LSE-C↑	5.653	7.555	4.138	4.716	6.060
	FID↓	21.572	11.847	9.437	23.158	6.734
	SSIM↑	0.5298	0.6072	0.9490	0.5526	0.9832
B	LSE-D↓	10.400	7.540	14.978	11.911	9.878
	LSE-C↑	5.234	6.066	0.238	2.135	4.963
	FID↓	19.983	13.120	10.234	19.315	7.065
	SSIM↑	0.6721	0.6049	0.9645	0.6238	0.9896
C	LSE-D↓	10.581	6.637	11.712	16.170	9.530
	LSE-C↑	5.122	8.951	3.322	0.06	6.141
	FID↓	19.311	11.154	9.677	19.46	6.498
	SSIM↑	0.350	0.5647	0.9316	0.4781	0.9744
D	LSE-D↓	10.005	6.546	11.804	11.444	9.091
	LSE-C↑	5.808	9.023	2.884	3.725	6.155
	FID↓	17.969	12.485	9.076	19.93	6.881
	SSIM↑	0.5682	0.5766	0.9439	0.4705	0.9875
E	LSE-D↓	12.506	6.571	11.713	14.339	9.831
	LSE-C↑	2.734	8.989	3.063	1.018	5.493
	FID↓	18.697	11.185	9.523	19.502	7.131
	SSIM↑	0.5716	0.6482	0.9237	0.4872	0.9867
F	LSE-D↓	9.567	6.343	9.953	11.167	8.817
	LSE-C↑	5.803	9.314	4.794	4.277	5.841
	FID↓	20.839	13.457	11.348	17.746	6.775
	SSIM↑	0.5321	0.7015	0.9577	0.4764	0.9894
G	LSE-D↓	9.687	6.013	13.332	8.831	8.880
	LSE-C↑	7.261	10.237	2.074	6.795	7.539
	FID↓	21.894	14.94	10.795	23.880	6.579
	SSIM↑	0.5794	0.6102	0.9102	0.7736	0.9866

1. The first and second places are marked in red and blue respectively.
2. The evaluation of LSE-D and LSE-C is conducted on audios from multi persons, while FID and SSIM test is driven by original audios to compare with original videos.

the center image frame are used as the paired image data to finally generate a 28 × 80 MFCC feature for each audio block.

All experiments were trained and tested on a single V100 using Pytorch. We compared our method with ATVG[3], Wav2lip[4], AudioDVP[2] and MakeIttalk[5] by testing their

Fig. 3. Comparison results of ATVG[3], Wav2lip[4], AudioDVP[2], MalkIttalk[5] and our method on video A, B and F driven by audio from multi persons.

Table 3. Ablation study on the test sets of videos.

Video	LSE-D	LSE-C	FID	SSIM
Baseline	11.955	2.930	10.007	0.940
Our A2E	9.578	6.343	9.264	0.952
High definition	11.834	3.427	7.449	0.976
Ours	9.237	6.027	6.804	0.985

driven performance on audio from multi persons. The comparison results among these methods are shown in Fig. 3. Our method generates more synchronized lip movements compared with the other four methods. The generated video can show more texture details of the face and even freckles on the F's face more clearly. Then, metrics LSE-D and LSE-C from [4] are adopted for quantitative evaluation of lip-syncing performance in the wild, and FID [4] and SSIM for image quality (see Table 2). The lip-sync performance of our method is comparable to Wav2Lip, and our method produces videos with the best image quality among these methods.

Table 3 shows our ablation study to prove our contribution to the improvemnt for talking face generation. We utilized the average score of the collected test dataset on AudioDVP[2] as a baseline to compare the scores of our A2E network and modification for high definition on LSE-D, LSE-C, FID and SSIM. Our improved A2E network and high-definition modification enhance the performance of speaking face video in terms of image quality and lip sync, respectively.

Table 4. User study results

Method	Average	→ Rating of realistic and definition →					synchronization
		1	2	3	4	5	'sync'
ATVG	1.98	32.1%	42.9%	19.6%	5.4%	0.0%	73.2%
Wav2Lip	3.23	0.0%	26.8%	30.4%	35.7%	7.1%	76.7%
AudioDVP	3.16	1.8%	16.1%	51.7%	25.0%	5.4%	26.7%
MakeIttalk	2.07	33.9%	30.4%	30.4%	5.3%	0.0%	32.1%
Ours	**4.45**	0.0%	0.0%	12.5%	30.4%	57.1%	83.9%

A user study was conducted to evaluate our method and state-of-the-art methods. We produced 35 short video clips of 768×768 size and 15s duration from the generated videos and randomly displayed to 20 anonymous participants. They were asked to evaluate the video clips from two perspectives: whether the generated talking face was realistic and high-definition and whether it was synchronized with the audio. Each clip was rated on a scale of 1 - 5 (5 for best image quality) by each participant. User study results in Table 4 show that our method can generate lip-synchronized realistic talking face video in most cases.

4. CONCLUSION

In this paper, we propose a new method to generate lip-synchronized talking face videos with high definition using only 3min video, achieving data-efficient training. Given any arbitrary audio input in the wild, it can drive the speech video generation of the target character in the test process. This method relieves the burden of target video collection and reduces the production cost for virtual reality application, which may lead to various potential applications. In the future work, we will focus on talking face generation based on target identity disentanglement.

5. REFERENCES

[1] "Synthesizing obama: Learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4CD, pp. 95.1–95.13, 2017.

[2] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, "Photorealistic audio-driven video portraits," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3457–3466, 2020.

[3] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7824–7833.

[4] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, New York, NY, USA, 2020, MM '20, p. 484–492, Association for Computing Machinery.

[5] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, "Makelttalk," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–15, Nov 2020.

[6] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo, "3d talking face with personalized pose dynamics," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[7] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," 2021.

[8] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo, "FACIAL: synthesizing dynamic talking face with implicit attribute learning," *CoRR*, vol. abs/2108.07938, 2021.

[9] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 285–295.

[11] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[12] Yudong Guo, juyong zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1294–1307, 2019.

[13] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

[14] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.