
Proceedings Paper:
Yang, Jingbo and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2022) Encoder Enabled GAN-based Video Generators. In: IEEE International Conference on Image Processing (ICIP). IEEE , Bordeaux, France , pp. 1841-1845.

https://doi.org/10.1109/ICIP46576.2022.9897233

ENCODER ENABLED GAN-BASED VIDEO GENERATORS

Jingbo Yang and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK
{jy1655, adrian.bors}@york.ac.uk

ABSTRACT

This research study proposes a compatible encoder-enabled video generating method. The encoder-enabled method adds an inference mechanism for enhancing the ability of Generative Adversarial Networks (GAN) based video generators. The proposed video generating method is called Encoding GAN3 (EncGAN3) and decomposes the video into two streams representing content and movement, respectively. The proposed model consists of three processing modules, representing Encoder, Generator and Discriminator, each trained separately, by considering its own loss function. EncGAN3 is shown to generate videos of high quality, according to both visual and numerical results.

Index Terms— Video generation, Generative Adversarial Network (GAN), Variational Autoencoder (VAE), Hybrid VAE-GAN model.

1. INTRODUCTION

The Generative Adversarial Network (GAN) [1] and the Variational Autoencoder (VAE) [2] represent the two main deep generative frameworks. GANs are able to generate sharp images but they are computationally expensive and sometimes result in unexpected artifacts during the generation. Meanwhile, VAEs are enabled by an inference mechanism with require comparably less computational cost and are more stable during training but tend to yield images which are blurry and of lower quality. The complementary characteristics of GAN and VAE consequentially resulted in the design of Hybrid VAE-GAN models aiming to overcome the weaknesses of both VAE and GAN models by combining their architectures [3, 4, 5, 6]. Current video generators are based on VAEs [7], GANs [8] as well as on their Hybrid architectures [9, 10].

In this paper, we propose a hybrid video generation model, employing an inference mechanism implemented by a compatible encoder for GAN-based generators aiming to provide realistic and high quality videos. The model named Encoding GAN3 (EncGAN3) uses an encoder for empowering GAN3 [8] with appropriate latent spaces in the content and movement spaces. The EncGAN3 model consists of a two-stream Encoder processing content and motion through two separate streams feeding a three-stream Generator ensuring the spatial, temporal and spatio-temporal consistency, as well as a two-stream Discriminator for image and video quality evaluation. The content and movement streams are processed separately and then fused at multiple scales into the main spatio-temporal consistency reconstruction stream eventually resulting in the generated video. The Encoder's dual stream architecture follows the success in processing separately content and motion for action recognition [11], as well as in other video generation methods [8, 12, 13, 14, 15, 16].

The encoder enables an appropriate inference mechanism to provide a representative latent space, instead of just using a random seed for the GAN generator. The useful information estimated from real data by inference benefits GAN models resulting in better video generation results. The proposed method has a wide application range and can be used in combination with most GAN-based video generators.

The contributions of this research study are as follows :

1. A new video generation approach by enabling GAN generators with video inference mechanisms.
2. A dual stream video generative model, in content and motion, namely Encoding GAN3 (EncGAN3).
3. Quantitative and qualitative results show the advantages of EncGAN3 with respect to the visual quality and diversity of generated videos.

2. RELATED WORKS

Hybrid VAE-GAN models in image generation attempt to combine the complementary characteristics of the GAN and VAE models for alleviating their weaknesses [3, 4, 5, 6]. VAE/GAN [4] uses the feature representation learned by a VAE to improve the data reconstruction produced by a GAN. However, when extending to video generation, these VAE-GAN designs lack scalability for processing temporal synchronization of moving objects and regions between consecutive frames.

Video generation methods initially followed the idea of generating sequences of consecutive temporally images. The Temporal Generative Adversarial Nets (TGAN) [13], Motion and Content GAN (MoCoGAN) [15] and Temporal Shift GAN (TS-GAN) [16] all use dual network architectures splitting the video generation process into image and sequence generators. VideoVAE [7] adds an additional time-processing

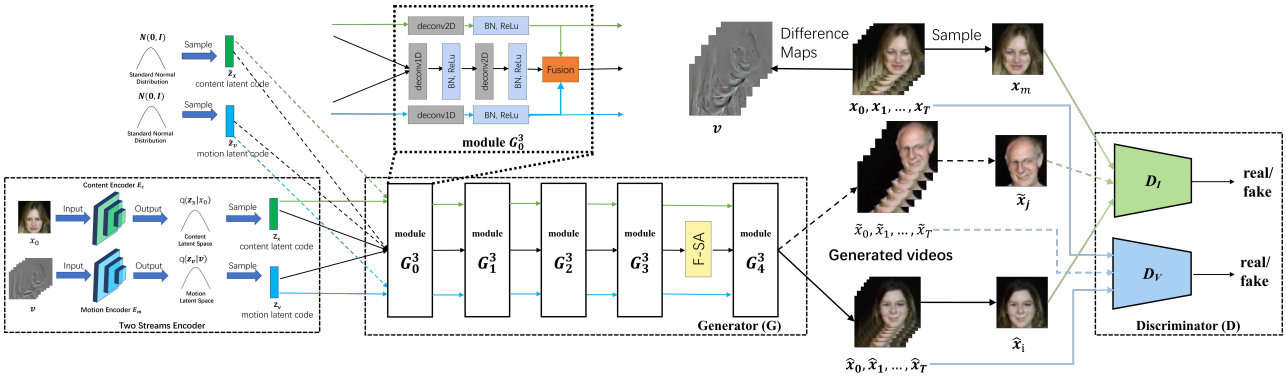


Fig. 1. The architecture of EncGAN3: two Encoders, a three-stream Generator and two Discriminators for deciding the content and movement information corresponding to the generated video.

module to the image encoded latent spaces for extending VAE-enabled image generation to generating sequences of images. TwoStreamVAN [10] is a model reconstructing video frames from a generated content frame and a set of difference maps between frames. G^{3}AN [8] generates videos using three streams representing motion, content and video reconstruction based on the GAN architecture.

3. THE ARCHITECTURE OF ENCGAN3

This paper proposes a new model, EncGAN3, for video generation. EncGAN3 enables multiple generative streams, as in G^{3}AN [8], enabled by a dual inference mechanism, as in [10]. The two mechanisms, of inference and generation, are matched in the latent spaces corresponding to the video content and its corresponding movement, represented by the temporal frame differences. The inference module encodes first frame of the input video to produce the content latent space. Meanwhile, the features of all difference maps between consecutive frames are encoded to produce the motion latent space. Separating the processing of motion and content corresponds to spatio-temporal decomposition of the information, which is widely used in video processing in applications for motion estimation, video compression, video classification and action recognition [11]. The motion stream encoder uses fully-connected layers to compress features from all difference maps, resulting in latent space codes, instead of using the Long Short Term Memory (LSTM) as in [10].

The architecture of EncGAN3, displayed in Fig. 1, consists of three modules: Encoder (Enc), Generator (G) and Discriminator (D). The video information is decomposed into content and movement before it is input to the two-stream Encoder resulting in content and motion latent codes, modeling probabilistic representations of video data. The Generator then transforms the two latent codes into three generation streams corresponding to the content, motion and video reconstruction. The video reconstruction stream combines the synchronized content and movement generation, resulting in a consistent sequence of video frames. Each of the four stacked modules, denoted as $\{G_i^3 | i = 0, \ldots, 4\}$ in G, fuses features from the three streams at different scales, such as the one of G_0^3 shown in the upper middle dashed box in Fig. 1. A factorized self-attention (F-SA) module [8] is placed between G_3^3 and G_4^3 aiming to improve the consistency of the generated video stream. The F-SA module consists of a temporal-wise followed by a spatial-wise self-attention module and enables the generator to utilize cues from all spatio-temporal features while modeling relationships between distinct regions. The processing pipeline ends with a two-stream Discriminator determining the realism of a randomly selected generated frame as well as for the whole generated video, separately.

In order to ensure a realistic representation of the generated video, the Generator input should match the prior distribution assumption of the latent space from the VAE. For ensuring this, we feed the Generator with the latent codes created by the motion and content encoders and also with noise data, sampled from a standard normal distribution, during the training, as shown by the dashed line of data flow in Fig. 1.

4. ENCGAN3 LOSS FUNCTIONS

The loss function is characteristic of hybrid VAE-GAN models [4, 6] used for image generation. Each of the three modules in EncGAN3, visualized from left to right in Fig. 1, has its own loss function and is trained individually.

First, the loss function of the two-stream encoder L_{Enc}, is defined as:

$$L_{Enc} = \sum_{i=1}^{N} \|\mathbf{x}_{i0} - \widehat{\mathbf{x}}_{i0}\| + \sum_{i=1}^{N}\sum_{j=0}^{T} \|\mathbf{x}_{ij} - \widehat{\mathbf{x}}_{ij}(\widehat{\mathbf{v}}_{ij_1}, \widehat{\mathbf{x}}_{i0})\| - D_{KL}(q_{\theta_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})||p(\mathbf{z}_{\mathbf{x}})) - D_{KL}(q_{\theta_{\mathbf{v}}}(\mathbf{z}_{\mathbf{v}}|\mathbf{v})||p(\mathbf{z}_{\mathbf{v}})) \quad (1)$$

where $\{\mathbf{x}_{ij}\}$ and $\{\widehat{\mathbf{x}}_{ij}\}$ are the j^{th} frame from the real i-th video and its corresponding reconstruction, respectively. $j \in \{0, \cdots, T\}$ while $j_1 \in \{1, \cdots, T\}$, where the latter represents the index of the difference map v_{ij_1}, calculated by subtracting consecutive frames. The Kullack-Leibler (KL) divergence D_{KL} enforces content and motion encoders to pro-

(a) EncGAN3, UvA dataset (b) EncGAN3, Weizmann dataset (c) EncGAN3, KTH dataset

Fig. 2. Video frames generated by EncGAN3 on UvA, Weizmann and KTH datasets from left to right. Every other row in (a) shows the difference maps used to represent the movement.

duce latent spaces $q_{\theta_\mathbf{x}}(\mathbf{z_x}|\mathbf{x})$ and $q_{\theta_\mathbf{v}}(\mathbf{z_v}|\mathbf{v})$ close to their assumed prior distributions $p(\mathbf{z_x})$ and $p(\mathbf{z_v})$, when optimizing parameters $\theta_\mathbf{x}$ and $\theta_\mathbf{v}$ for content and motion encoders. Both prior distributions are set as standard Normal distributions, $\mathcal{N}(0, \mathbf{I})$. The two-stream encoders trained together for better performance, as in L_{Enc}, rather than separate.

Secondly, the loss function of the Generator L_G contains both VAE and GAN loss components:

$$\begin{aligned} L_G =& \mathbb{E}_{\widehat{\mathbf{x}}_n \sim G(\mathbf{z_x}, \mathbf{z_v})} \log[D(\widehat{\mathbf{x}}_n)] + \mathbb{E}_{\tilde{\mathbf{x}}_n \sim G(\tilde{\mathbf{z}}_\mathbf{x}, \tilde{\mathbf{z}}_\mathbf{v})} \log[D(\tilde{\mathbf{x}}_n)] \\ &+ \mathbb{E}_{\mathbf{z_x} \sim q_{\theta_\mathbf{x}}(\mathbf{z_x}|\mathbf{x}), \mathbf{z_v} \sim q_{\theta_\mathbf{v}}(\mathbf{z_v}|\mathbf{v})} \log[D(G(\mathbf{z_x}, \mathbf{z_v}))] \\ &+ \mathbb{E}_{\tilde{\mathbf{z}}_\mathbf{x} \sim \mathcal{N}(0,\mathbf{I}), \tilde{\mathbf{z}}_\mathbf{v} \sim \mathcal{N}(0,\mathbf{I})} \log[D(G(\tilde{\mathbf{z}}_\mathbf{x}, \tilde{\mathbf{z}}_\mathbf{v}))] \\ &+ \sum_{i=1}^{N} \|\mathbf{x}_{i0} - \widehat{\mathbf{x}}_{i0}\| + \sum_{i=1}^{N} \sum_{j=0}^{T} \|\mathbf{x}_{ij} - \widehat{\mathbf{x}}_{ij}(\widehat{\mathbf{v}}_{ij_1}, \widehat{\mathbf{x}}_{i0})\| \end{aligned} \tag{2}$$

where $\tilde{\mathbf{x}}_n$ is a randomly picked frame from the video reconstructed from the content and motion noises $\tilde{\mathbf{z}}_\mathbf{x}$ and $\tilde{\mathbf{z}}_\mathbf{v}$. Both random samples are sampled from the assumed Normal prior distribution of the latent space, instead of being inferred from the latent spaces provided by the encoder. By using random variables $\tilde{\mathbf{z}}_\mathbf{x}$ and $\tilde{\mathbf{z}}_\mathbf{v}$ aims to enforce that G learns to reconstruct well based on these inputs from prior distribution because L_G does not have a KL divergence component. The last two terms correspond to the reconstruction errors from the VAE loss, which are also part of L_{Enc} from Eq. (1), while the other terms correspond to the GAN loss.

Thirdly, the loss function of the two-stream Discriminator is an adversarial loss. Each of the three streams has its own Discriminator while all three are trained in parallel. The loss function of the image-stream Discriminator L_{D_I} is :

$$\begin{aligned} L_{D_I} =& \mathbb{E}_{\mathbf{x}_n \sim p(\mathbf{x})} \log[D(\mathbf{x}_n)] \\ &+ \mathbb{E}_{\widehat{\mathbf{x}}_n \sim G(\mathbf{z_x}, \mathbf{z_v})} \log[1 - D(\widehat{\mathbf{x}}_n)] \\ &+ \mathbb{E}_{\tilde{\mathbf{x}}_n \sim G(\tilde{\mathbf{z}}_\mathbf{x}, \tilde{\mathbf{z}}_\mathbf{v})} \log[1 - D(\tilde{\mathbf{x}}_n)] \end{aligned} \tag{3}$$

where $\mathbf{x}_n \sim p(\mathbf{x})$ is a frame sampled from the real video, $\widehat{\mathbf{x}}_n$ is from the video generated by latent codes and $\tilde{\mathbf{x}}_n$ is the one generated using $\mathcal{N}(0, \mathbf{I})$.

Eventually, there is the loss function L_{D_V} for the video-stream Discriminator :

$$\begin{aligned} L_{D_V} =& \mathbb{E}_{\mathbf{x}_{0:T} \sim p(\mathbf{x}_{0:T})} \log[D(\mathbf{x}_{0:T})] \\ &+ \mathbb{E}_{\tilde{\mathbf{z}}_\mathbf{x} \sim \mathcal{N}(0,\mathbf{I}), \tilde{\mathbf{z}}_\mathbf{v} \sim \mathcal{N}(0,\mathbf{I})} \log[1 - D(G(\tilde{\mathbf{z}}_\mathbf{x}, \tilde{\mathbf{z}}_\mathbf{v})] \\ &+ \mathbb{E}_{\widehat{\mathbf{x}}_{0:T} \sim p(\widehat{\mathbf{x}}_{0:T})} \log[1 - D(\widehat{\mathbf{x}}_{0:T})], \end{aligned} \tag{4}$$

where $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$ and $\widehat{\mathbf{x}}_{0:T} = \{\widehat{\mathbf{x}}_0, \ldots, \widehat{\mathbf{x}}_T\}$ represent the real videos and their reconstructions, while $p(\mathbf{x}_{0:T})$ and $p(\widehat{\mathbf{x}}_{0:T})$ are their probabilistic representations.

	UvA FID↓	Weizmann FID↓	KTH FID↓
VGAN*	235.01	158.04	-
TGAN*	216.41	99.85	-
MoCoGAN*	197.32	92.18	-
G^{3}AN	91.77	98.27	111.99
EncGAN3	**87.63**	**83.35**	**72.59**

Table 1. FID↓ implies that lower FID means better visual quality and spatio-temporal consistency. "*" results are referred from [8].

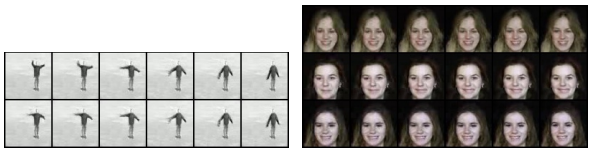
	IS↑	H(y)↑	H(y|x)↓	Dataset
G^{3}AN	85.44	6.041	1.593	UvA
	25.54	3.924	0.684	Weizmann
	24.19	4.538	1.352	KTH
EncGAN3	571.29	6.499	0.151	UvA
	42.60	3.959	0.207	Weizmann
	50.48	4.812	0.891	KTH

Table 2. IS and its components for EncGAN3 and G^{3}AN.

During the training, the Discriminator D is firstly updated by optimizing L_{D_I} and L_{D_V} using equations (3) and (4), then the Encoder using L_{Enc} from Eq. (1), and eventually we rerun the model with the optimized Discriminator and Encoder on the same video and considering $(\tilde{\mathbf{z}}_\mathbf{x}, \tilde{\mathbf{z}}_\mathbf{v})$ as input to update the Generator L_G according to Eq. (2)

5. EXPERIMENTS

We generate video sequences of 16 frames ($T = 15$) and of resolution 64×64 pixels, using EncGAN3, trained on the PyTorch deep learning platform, considering the learning rates



(a) z_{c_1},z_{m_1} (top) and z_{c_1},z_{m_2} (bottom)

(b) z_{c_3},z_{m_3} (top), z_{c_4},z_{m_4} (middle) and their sum (bottom)

Fig. 3. Manipulating latent codes to generate related videos.

of 2×10^{-4}. Frames generated by EncGAN3 after training on UvA-NEMO [17], Weizmann [18] and KTH [19] datasets are shown in Figures 2-a, b and c, respectively.

We evaluate EncGAN3 model performance using the Fréchet Inception Distance (FID) [20] and Inception Score (IS) [21]. Lower FID means better visual quality and spatio-temporal consistency while higher IS represents better visual quality and diversity. Inter-Entropy H(y) and Intra-Entropy H(y|x) are components of the IS [7]. A higher H(y) indicates better generated video diversity while lower H(y|x) means better visual quality of generated videos. In Table 1, we compare the FID for videos generated by EncGAN3 with G^{3}AN [8], VGAN [12], TGAN [13] and MoCoGAN [15]. Table 2 presents IS and its corresponding Inter-Entropy and Intra-Entropy terms H(y) and H(y|x), where it can be observed that EncGAN3 consistently achieves best results.

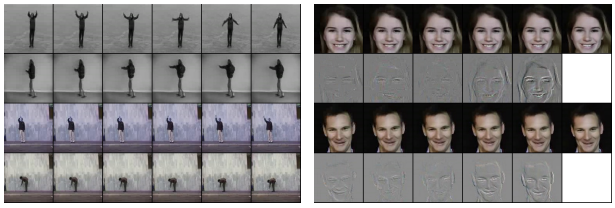
	F-SA	UvA	Weizmann	KTH
G^{3}AN	no	95.47	89.98	79.36
G^{3}AN	yes	91.77	88.08	75.38
EncGAN3(G)	no	89.46	88.00	**62.53**
EncGAN3(G)	yes	93.65	102.36	83.51
EncGAN3(Enc+G)	no	87.52	**82.43**	73.79
EncGAN3(Enc+G)	yes	**87.16**	83.36	72.59

Table 3. FID score when excluding some modules.

Architecture	UvA		Weizmann		KTH	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
no G_S,G_T	95.500	63.926	101.638	2.244	73.220	2.867
no G_S	88.058	133.352	89.004	7.020	75.309	3.853
no G_T	90.713	537.852	97.554	5.564	74.963	4.966
EncGAN3	87.16	571.29	78.935	8.906	70.448	5.986

Table 4. Contributions of the spatial G_S and temporal G_T streams in the Generator G.

We perform an ablation study of specific components of the EncGAN3. From Table 3 we observe that the presence of the Encoder has a more substantial improvement than the F-SA module, which only brings a small improvement. In Table 3 EncGAN3 (G) would generate videos using only the Generator while EncGAN3 (Enc+G) would use both Encoder and Generator. We also evaluate the dual stream processing by using generators G_S and G_T for content and movement streams. From Table 4 the two auxiliary streams G_S and G_T are necessary and clearly improving the model performance.



(a) KTH and Weizmann (b) UvA-NEMO

Fig. 4. Generated frames of 128×128 pixel resolution.

Fig. 5. Generated videos with two moving objects.

We are also exploring the relationship between latent codes and the generated frames as shown in Fig. 3. Frames from the first and second rows of Fig. 3-a are generated from the same content latent code but with different motion latent codes based on videos from KTH dataset showing the same person performing different movements. In Fig. 3-b, frames from the third row are generated from the latent codes representing the sum of those used for the first and second rows. It can be observed that the person in the third row inherits facial properties from those shown in the rows above.

We test the EncGAN3 performance on generating videos with the resolution of 128×128 pixels with the results shown in Fig. 4, by adding a further G^3 module to the previous structure used for generating 64×64 pixels video frames. The top two rows in Fig. 4-a show frames from KTH while the bottom two rows from Weizmann. Fig. 4-b are from the UvA dataset and the difference maps from under the face images in (b) indicate the ability of EncGAN3 to generate subtle facial expression movement as in micro-expressions [22]. Unlike any other video generation method, EncGAN3 is also able to generate complex videos with multiple moving objects, as shown in Fig. 5 after being trained on the Weizmann dataset. Fig. 5 shows two persons doing either similar or different movements at the same time in 128×128 pixels videos.

6. CONCLUSION

In this research study, we propose to enable GAN-based video generation models with inference mechanisms by embedding an encoder, instead of the random generator. We consider a dual stream generation process, for content and movement streams, in both the encoder and generator within the EncGAN3 model. The proposed model is shown to generate realistic video sequences of resolutions up to 128×128, with characteristics that can be controlled through latent codes and even displaying multiple moving objects.

7. REFERENCES

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, p. 2672–2680. 1

[2] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:11312.6114*, 2014. 1

[3] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," in *arXiv preprint arXiv:1511.05644*, 2015. 1

[4] Anders Larsen, Søren Sønderby, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2016, pp. 1558–1566. 1, 2

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2745–2754. 1

[6] Fei Ye and Adrian G. Bors, "Learning joint latent representations based on information maximization," *Information Sciences*, vol. 567, no. 8, pp. 216–236, 2021. 1, 2

[7] J. He, A. Lehrmann, J. Marino, G. Mori, and L. Sigal, "Probabilistic video generation using holistic attribute control," in *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 11209*, 2018, pp. 466–483. 1, 4

[8] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva, "G3AN: Disentangling Appearance and Motion for Video Generation," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5264–5273. 1, 2, 3, 4

[9] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva, "ImaGINator: Conditional Spatio-Temporal GAN for Video Generation," in *Proc. IEEE/CVF Winter Conf. on Applic. of Computer Vision (WACV)*, 2020, pp. 1160–1169. 1

[10] Ximeng Sun, Huijuan Xu, and Kate Saenko, "TwoStreamVAN: Improving motion modeling in video generation," in *Proc. IEEE/CVF Winter Applic. in Computer Vison (WACV)*, 2020, pp. 2744–2753. 1, 2

[11] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308. 1, 2

[12] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba, "Generating Videos with Scene Dynamics," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 613–621. 1, 4

[13] Masaki Saito, Eiichi Matsumoto, and Shunta Saito, "Temporal Generative Adversarial Nets With Singular Value Clipping," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2830–2839. 1, 4

[14] Aidan Clark, Jeff Donahue, and Karen Simonyan, "Adversarial video generation on complex datasets," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1907.06571*, 2019. 1

[15] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1526–1535. 1, 4

[16] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox, "Temporal shift gan for large scale video generation," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, January 2021, pp. 3179–3188. 1

[17] Hamdi Dibeklioğlu, Albert A. Salah, and Theo Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 7574*, 2012, pp. 525–538. 4

[18] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and M. Basri, "Actions as space-time shapes," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2005, pp. 1395–1402. 4

[19] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. on Pattern Recog. (ICPR)*, 2004, vol. 3, pp. 32 – 36. 4

[20] 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," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6626–6637. 4

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, vol. 29. 4

[22] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman, "Imaging facial physiology for the detection of deceit," *International Journal on Computer Vision*, vol. 71, no. 2, pp. 197–214, 2007. 4