ENABLING THE ENCODER-EMPOWERED GAN-BASED VIDEO GENERATORS FOR LONG VIDEO GENERATION

Jingbo Yang and Adrian G. Bors

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

2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222725

ABSTRACT

Despite the remarkable progress in the video generation field, generating videos of longer-term remains challenging due to the challenge of sustaining the temporal consistency and continuity in the resulting synthesized movement while ensuring realism. In this paper, we propose a recall mechanism for enabling an encoder-empowered short-term video generator to produce long-term videos. This mechanism connects smoothly short video clips by modeling their temporal connections. We propose the Recall Encoder-GAN3 (REncGAN3), which enables an Encoder-based Generative Adversarial Network (GAN) to connect short generated video clips into longer sequences of hundreds of frames. The recall mechanism, defined through a loss function, enables an appropriate plasticity-continuity balance in the resulting long video stream. The proposed long-term video generation method ensures the generation of several hundred frames displaying consistent movement, which is non-repetitive while the computational memory costs are similar to those of short video generation models.

Index Terms— VAE-GAN architectures, Long-term video generation.

1. INTRODUCTION

Video generation methods aim to use the relationship between frames for modeling movement in the scene or by reducing the redundancy between consecutive frames. Generative Adversarial Networks (GANs) [1, 2, 3, 4], Variational Autoencoder (VAEs) [5, 6, 7, 8, 9, 10] or their extensions [11, 12, 13, 14, 15, 16, 17] have been used as video generative models. Most existing video generation methods are shown to generate about 16 frames and most would struggle to generate realistic videos of more than a hundred frames. In this study, we called models that generate less than 100 frames as short-term video generation methods [18, 19, 20] to distinguish them from the recently proposed long-term video generation methods [21, 22, 23]. The general synthesis of each frame relies on modeling temporal dynamics and coherence of the entire sequence, significantly increasing computational complexity and memory costs for generating longer sequences.

In this paper, we propose extending existing short-term video generation approaches to generating long-term videos by using a recall mechanism. We consider that we have short videos produced by a hybrid VAE-GAN video generation model, namely Encoding GAN3 (EncGAN3) proposed in [13]. EncGAN3 employs an inference mechanism implemented by a dual stream encoder feeding data to a GAN-based generator disentangling movement from content in a dual generation stream [19]. The dual stream consists of reconstructing content, represented as an image frame, and movement, modeled by differences between consecutive video frames. We propose the Recall EncGAN3 (REncGAN3) which applies the proposed recall mechanism to utilize the VAE-GAN hybrid short-term video generator from EncGAN3 [13], for generating long-term video sequences. In REncGAN3, the recall mechanism utilizes the inference properties of the encoder to enable the modeling of temporal relationships between consecutive clips instead of within the frames of the entire sequence, saving representation cost. Hence, the generator in REncGAN3 produces connected clips instead of a whole sequence made up of individual frames. REncGAN3 trains the encoder and generator jointly for improving the generation of long-term videos, unlike in EncGAN3 where these are trained separately. Meanwhile, REncGAN3 requires identical GPU memory requirements as EncGAN3 while is able to generate long-term videos of hundreds of frames with good temporal consistency.

The following contributions are brought in this paper : a) We propose a new recall-based model REncGAN3 for long-term video generation modeling temporal relationships between consecutive video clips; b) Quantitative and qualitative results for REncGAN3 showing good visual quality and displaying spatial-temporal consistency and stability of generated long-term videos.

2. THE ARCHITECTURE OF RENCGAN3

Th architecture of the proposed Recall EncGAN3 (REncGAN3), designed for generating long-term videos is shown in Fig. 1. REncGAN3 relies upon EncGAN3 [13] to generate short video clips and then through training enforces the temporal connectivity between the generated clips through the recall mechanism. EncGAN3 [13] is a VAE-GAN hybrid method enabling a GAN-based video generator with encoders

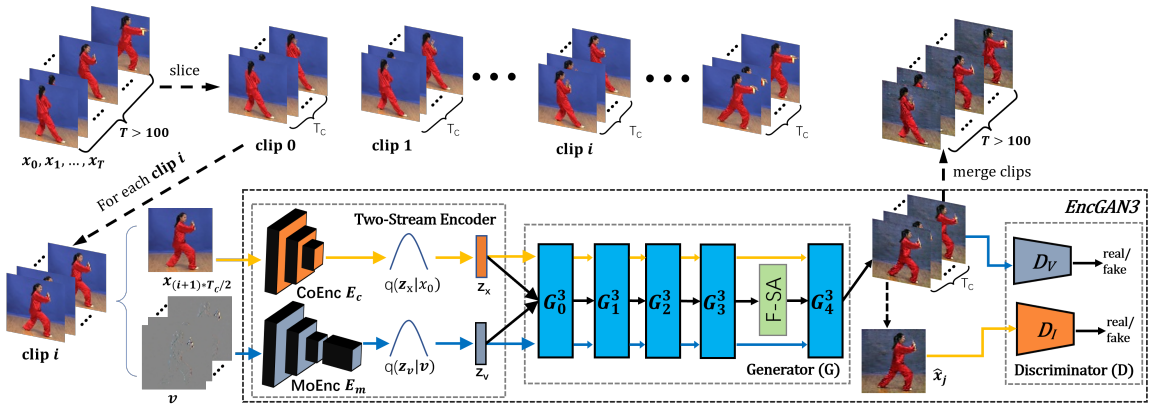

Fig. 1. The architecture of REncGAN3 continuously interlinks successive video clips for generating long video sequences.

in order to infer useful information for generating better video results. In the processing order, EncGAN3 consists of a two-stream Encoder, a three-stream Generator and a two-stream Discriminator [13]. Two separate generating streams, one for content and another for motion, are propagating through two encoders, E_c and E_m, for each of the two streams, and then enabling the generator G to generate appropriate content and movement data, as shown in the lower part of Fig. 1. As shown in the upper part of Fig. 1, each long video sequence is separated into several overlapping video clips, of T_c-frames each (in the experiments $T_c = 16$). Through the recall mechanism, a number of overlapping frames is considered between each two consecutive clips defined by a reference frame $\mathbf{x}_r$. The reference frame $\mathbf{x}_{i,T_c-r}$ of clip i corresponds to the frame $\mathbf{x}_{i+1,0}$ of next clip. Meanwhile, the reference frame $\mathbf{x}_{i+1,r}$ of clip $i+1$ is the same as the last frame of its previous clip i. The difference of indexes $T_c - r$ (r for the initial video segment) requires training two EncGAN3 modules to recursively connect all clips in a long video. We consider $T_c - r = r$, resulting in $r = T_c/2$, enabling each generated clip to half-overlap with both its previous and next clips, thus ensuring the continuity and consistency of the movement.

3. LOSS FUNCTIONS OF RENCGAN3

In REncGAN3, we expand the EncGAN3 model for generating probabilistic dependencies by enforcing the continuity between successive small video clips leading to longer videos. A long video $\widehat{\mathbf{y}}_{1:T}$ is created by recursively connecting pairs of shorter video clips $\widehat{\mathbf{x}}_{i,1:T_c}$, $i = 1,\ldots,N$ for an entire sequence of N generated video clips, $T \gg T_c$. The continuity between consecutive segments of the long video is ensured through a Markov chain by considering a reference frame $\widehat{\mathbf{x}}_{i,T_c/2}$ from one video clip i to the next one $i+1$, $i = 1,\ldots,N-1$ for linking successive video clips :

$$p(\widehat{\mathbf{y}}) = \prod_{i=1}^{N} p(\widehat{\mathbf{x}}_{i,1:T_c}) = p(\widehat{\mathbf{x}}_{1,1:T_c}) \prod_{i=2}^{N} p(\widehat{\mathbf{x}}_{i,1:T_c} \mid \widehat{\mathbf{x}}_{i-1,1:T_c})$$

$$\approx p(\widehat{\mathbf{x}}_{1,1:T_c}) \prod_{i=2}^{N} p(\widehat{\mathbf{x}}_{i,1:T_c} \mid \widehat{\mathbf{x}}_{i-1,r}), \quad (1)$$

where we consider $\widehat{\mathbf{x}}_{i,1:T_c}$ for the short video sequences of length T_c for $i = 1,\ldots,N_c$, while $\widehat{\mathbf{x}}_{i-1,r}$, represent the reference frame from $(i-1)$th video clip.

The loss function for training together the Encoder and Generator in REncGAN3 is given by :

$$\begin{aligned} L_{EncG} = & \sum_{m=1}^{N_L}\sum_{i=1}^{N_C} \|\mathbf{x}_{m,i,r} - \widehat{\mathbf{x}}_{m,i,r}\| \\ & + \sum_{m=1}^{N_L}\sum_{i=1}^{N_C}\sum_{j=1}^{T_c-1} \|\mathbf{x}_{m,i,j} - \widehat{\mathbf{x}}_{m,i,j}(\widehat{\mathbf{v}}_{m,i,j}, \widehat{\mathbf{x}}_{m,i,r})\| \\ & + D_{KL}(q_{\theta_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})||p(\mathbf{z}_{\mathbf{x}})) \\ & + D_{KL}(q_{\theta_{\mathbf{v}}}(\mathbf{z}_{\mathbf{v}}|\mathbf{v})||p(\mathbf{z}_{\mathbf{v}})) \\ & - \mathbb{E}_{\mathbf{z}_{\mathbf{x}}\sim q_{\theta_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x}),\mathbf{z}_{\mathbf{v}}\sim q_{\theta_{\mathbf{v}}}(\mathbf{z}_{\mathbf{v}}|\mathbf{v})} \log[D(G(\mathbf{z}_{\mathbf{x}},\mathbf{z}_{\mathbf{v}}))] \\ & - \mathbb{E}_{\widehat{\mathbf{x}}_n\sim G(\mathbf{z}_{\mathbf{x}},\mathbf{z}_{\mathbf{v}})} \log[D(\widehat{\mathbf{x}}_n)] \end{aligned} \quad (2)$$

where we have N_L long-term videos, with each split into N_C overlapped clips, each clip containing T_c frames. Each $\{\mathbf{x}_{m,i,j}\}_{j=0}^{T_c-1}$ represents an image frame while $\widehat{\mathbf{x}}_{m,i,j}$ is its reconstruction. $\{\widehat{\mathbf{v}}_{m,i,j}\}_{j=1}^{T_c-1}$ represents the reconstruction of the movement, as the differences between consecutive frames, associated with the frame j form clip i from the long-term video m. Meanwhile, $\{\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{v}}\}$, represent the latent spaces of the content and movement, modeled by the encoders E_c and E_m, respectively. The loss function for REncGAN3 from Eq. (2) trains the Encoder and Generator together to integrate better their latent spaces, which is different from EncGAN3 in [13], where these are trained separately.

The recall mechanism uses the two-stream encoder of EncGAN3 to relate the information from the ending of a video clip with that from the beginning of the next video clip, for enforcing the continuity and consistency in the long-term video sequence. Thus, the loss function reconstruction error term enforces the generator to learn the connecting information for generating consistent and coherent video components from individual video clips. The image reconstruction error term restricts the reference frame of the generated clip to be close to the input of the next video clip while the video reconstruction error term restricts the other frames of the

generated clip to be coherent with the reference frame. As the input clips are overlapped through their reference frame, the clips represent continuously connected consecutive video segments of the same long-term generated video. The random generation property of the GAN provides for the diversity of generated clips, resulting in diverse long videos.

For each clip, the frame reconstructions $\{\widehat{\mathbf{x}}_{ij}\}_{j=0}^{T_c-1}$ are calculated recursively using the reconstructed reference frame $\widehat{\mathbf{x}}_{i,r}$ and frame differences reconstructions $\{\widehat{\mathbf{v}}_{ij}\}_{j=1}^{T_c-1}$ as :

$$\begin{aligned} \widehat{\mathbf{x}}_{i,j-1} &= \widehat{\mathbf{x}}_{i,j} \ominus \widehat{\mathbf{v}}_{i,j},\ j = 1, \ldots, r \\ \widehat{\mathbf{x}}_{i,j} &= \widehat{\mathbf{x}}_{i,j-1} \oplus \widehat{\mathbf{v}}_{i,j},\ j = r+1, \ldots, T_c - 1 \end{aligned} \tag{3}$$

where $i = 1, \ldots, N_c$ and the index of the reference frame is considered as $r = T_c/2$ in REncGAN, while $\ominus$ and $\oplus$ mean pixel-wise addition and subtraction, respectively. By considering $r = T_c/2$, we ensure that consistency of the current video clip in equal proportions with the next and previous video clips ensuring overall video consistency and continuity.

For the two-stream Discriminator, we have two loss functions L_{D_I} and L_{D_V}, for deciding how realistic are the image and video streams. The Discriminator of each stream is trained independently and both Discriminators are optimized in parallel, similar with [11, 13, 18, 19]. The loss function of the image-stream Discriminator L_{D_I} is given by :

$$\begin{aligned} L_{D_I} = &-\mathbb{E}_{\mathbf{x}_n \sim p(\mathbf{x})} \log[D(\mathbf{x}_n)] \\ &- \mathbb{E}_{\widehat{\mathbf{x}}_n \sim G(\mathbf{z_x}, \mathbf{z_v})} \log[1 - D(\widehat{\mathbf{x}}_n)] \end{aligned} \tag{4}$$

where $\mathbf{x}_n$ is a frame sampled from the real video clip and $\widehat{\mathbf{x}}_n$ is from the video generated by latent codes.

The video-stream Discriminator L_{D_V} loss function is :

$$\begin{aligned} L_{D_V} = &-\mathbb{E}_{\mathbf{x}_{0:T_c} \sim p(\mathbf{x}_{0:T_c})} \log[D(\mathbf{x}_{0:T_c})] \\ &- \mathbb{E}_{\widehat{\mathbf{x}}_{0:T_c} \sim p(\widehat{\mathbf{x}}_{0:T_c})} \log[1 - D(\widehat{\mathbf{x}}_{0:T_c})] \end{aligned} \tag{5}$$

where $\mathbf{x}_{0:T_c} = \{\mathbf{x}_{ij}\}_{j=0}^{T_c}$ and $\widehat{\mathbf{x}}_{0:T_c} = \{\widehat{\mathbf{x}}_{ij}\}_{j=0}^{T_c}$ represent the real videos and their generations, while $p(\mathbf{x}_{0:T_c})$ and $p(\widehat{\mathbf{x}}_{0:T_c})$ are their probabilities.

During the training, first the Discriminator is updated by optimizing L_{D_I} and L_{D_V} using Eq. (4) and (5), then the Encoder and Generator by L_{EncG}, according to Eq. (2).

4. EXPERIMENTS

We train REncGAN3 using the loss functions described in Section 3 on Tai-Chi-HD (Taichi) [24] dataset to generate long video sequences with video lengths of hundreds of frames. The video lengths of generated long-term videos depend on the lengths of input video data. Initially, we had generated video clips of length $T_c = 16$, using the training explained in Section 3, and then by considering overlapping of 50%, *i.e.* $T_c/2 = 8$ frames overlapping between consecutive video clips, enabling the recall mechanism for generating long-term videos. The training of REncGAN3 is implemented using the ADAM optimizer [25] with the exponential decay rate of first-order and second-order moment estimation of β_1=0.5 and β_2=0.999, while considering a learning rate of $2e^{-4}$ for the loss functions when training all modules: Discriminator, Encoder and Generator.

Frames sampled from a long-term video generated following the training on the Tai-Chi-HD dataset are shown in the first row of images from Fig. 2, while underneath on the second row we provide the frames when adapting the loss function of EncGAN3 [13] for training REncGAN3 processing configuration for the long-term video sequence and name this approach as REncGAN3 (Enc_G). The main difference is that in REncGAN3 (Enc_G) the Encoder and Generator are trained separately instead of jointly end-to-end as proposed by using L_{EncG} from Eq. (2) for REncGAN3. The bottom two rows from Fig. 2 show frames from the videos generated by DIGAN [22] and TATS [23], respectively. It can be observed that the frames generated by REncGAN3 show temporally consistency and continuity, while the results by DIGAN [22] cannot maintain the consistency of the representation in many frames, displaying blurred features, while TATS [23] produces repeated movements which are not consistent with the Tai-Chi action. REncGAN3 produces Taichi movements which have the appropriate speed, fitting the original movements, while the other methods generate videos displaying rather quick movements. Meanwhile, the frames generated by REncGAN3 show better visual quality than REncGAN3 (Enc_G). Moreover, REncGAN3 is able to generate videos of various lengths by means of a simple recall mechanism connecting short-term clips.

In the following we evaluate the Fréchet Inception Distance (FID) [26] on sequences of 16-frame clips which are cut sequentially from the long-term videos generated by REncGAN3. Lower FID values indicate high visual quality and spatial-temporal consistency of generated videos. The video FID values for 3 long-term videos generated after training on the TaiChi dataset together with their average are provided in Fig. 3. Two of the generated videos display consistency with good FID scores, while the one indicated in red and labeled as '560f', displays more complex movement with some segments characterized by high FID scores. The recall mechanism in REncGAN3 merges short-term clips forming a long-term video instead of generating the whole long video frame by frame, displaying quality consistency, as shown by the FID results from Fig. 3. Moreover, REncGAN3 requires for generating long-term videos the same amount of GPU memory as EncGAN3, which can only generate short sequences.

For the quantitative evaluation, we consider the generation of short-term video clips of 16 frames each. The results for FID are provided in Table 1, where "*" indicates that results are referred from [13, 19]. REncGAN3 uses the loss function from Eq. (2) for the end-to-end training of the Encoder and Generator modules, while EncGAN3, proposed in

Fig. 2. Each row from top to bottom show frames from videos generated by REncGAN3, EncGAN3 (Enc_G), DIGAN [22] and TATS [23], respectively. Videos in each row have lengths of 400, 424, 1024 and 1024 frames with the same resolution of 128×128. To illustrate the long sequence within limited space, frames in each row are sampled with steps of 8 frames from the sequences 0 to 130 (left), 130 to 260 (middle) and 260 to 400 (right).

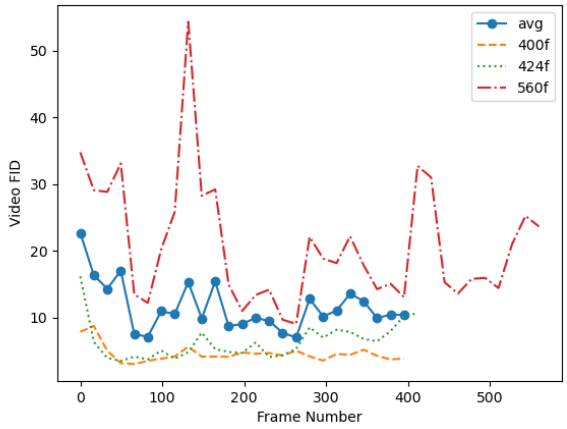

Fig. 3. **Quality degradation.** Video FID evaluated for successions of non-overlapping 16-frame clips split from long-term videos generated by REncGAN3, after being trained on the Taichi dataset for lines labeled with 'f' to show video length while "avg" is averaging.

[13], trains them separately. We also consider the Inception Score (IS) [27] for videos [11], called video IS. Higher video IS indicates better visual quality and diversity. The Inter-Entropy H(y) and Intra-Entropy H(y|x) [6] are components of the video IS that measure the visual quality and diversity of generated videos, respectively. The results from Table 2 compare REncGAN3 with EncGAN3. REncGAN3 provides significantly better spatial-temporal consistency (lower video FID score), though with worse diversity (lower H(y)) and visual quality (higher H(y|x)) than EncGAN3 on the face expression UvA, Human Action Weizmann and KTH datasets. These results show the benefits to the spatial-temporal consistency of generated videos when trained with tighter connections between the Encoder and Generator through using the loss function from Eq. (2), although of a lower diversity. When training on the bigger and more complex UCF101 dataset which contains videos with variations of facial expressions and human actions, REncGAN3 achieves significantly better visual quality but provides worse diversity and spatial-temporal consistency than EncGAN3.

	UvA FID↓	Weizmann FID↓	KTH FID↓	UCF101 FID↓
VGAN*	235.01	158.04	-	115.06
TGAN*	216.41	99.85	-	110.58
MoCoGAN*	197.32	92.18	-	104.14
G^3AN*	91.77	98.27	111.99	108.36
EncGAN3	87.63	83.35	72.59	**91.18**
REncGAN3	**73.14**	**70.91**	**66.97**	95.87

Table 1. Video FID results, where "*" indicate that results are referred from [19, 13].

	IS↑	H(y)↑	H(y|x)↓	Dataset
EncGAN3	571.29	6.499	0.151	UvA
	42.60	3.959	0.207	Weizmann
	50.48	4.812	0.891	KTH
	33.87	6.699	3.177	UCF101
REncGAN3	87.007	4.656	0.190	UvA
	35.329	3.804	0.239	Weizmann
	11.477	4.087	1.647	KTH
	57.121	5.827	1.782	UCF101

Table 2. IS and its components, where ↑ indicates that higher values are better, while ↓ shows that lower values are better.

5. CONCLUSION

In this paper, we propose generating long-term video sequences by learning temporal relationships between short video clips. A recall mechanism, relying on an encoder-based inference mechanism, is used for enabling a short-term video generator to learn temporal relationships between consecutive clips. This mechanism is applied to a hybrid Encoder-GAN short-term video generator, resulting in the recall EncGAN3 (REncGAN3). In REncGAN3, the recall mechanism utilizes the inference mechanism in EncGAN3 to enable the model to learn not only the temporal relationships between frames within the 16-frame clip but also the temporal relationships between pairs of short video clips. The proposed REncGAN3 generates videos with hundreds of consistent frames displaying continuity and consistency.

6. REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2014, p. 2672–2680.

[2] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating Videos with Scene Dynamics," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 613–621.

[3] 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.

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

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

[6] 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.

[7] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci, "Playable video generation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10061–10070.

[8] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer, "ipoke: Poking a still image for controlled stochastic video synthesis," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14707–14717.

[9] Lingyun Song, Jun Liu, Buyue Qian, and Yihe Chen, "Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[10] Lingyun Song, Jun Liu, Buyue Qian, Mingxuan Sun, Kuan Yang, Meng Sun, and Samar Abbas, "A deep multi-modal cnn for multi-instance multi-label image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6025–6038, 2018.

[11] 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.

[12] 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.

[13] Jingbo Yang and Adrian G. Bors, "Encoder enabled GAN-based video generators," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 2022, pp. 1841–1845.

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

[15] Fei Ye and Adrian G. Bors, "Learning latent representations across multiple data domains using lifelong VAEGAN," in *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 12365*, 2020, pp. 777–795.

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

[17] A. Larsen, S. Sønderby, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2016, pp. 1558–1566.

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

[19] 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.

[20] Pierfrancesco Ardino, Marco De Nadai, Bruno Lepri, Elisa Ricci, and Stéphane Lathuilière, "Click to move: Controlling video generation with sparse motion," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14749–14758.

[21] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3626–3636.

[22] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin, "Generating videos with dynamics-aware implicit generative adversarial networks," in *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2202.10571*, 2022.

[23] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh, "Long video generation with time-agnostic vqgan and time-sensitive transformer," *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 13677*, pp. 102–118, 2022.

[24] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1412.6980*, 2015.

[26] 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 (NeurIPS)*, 2017, pp. 6629–6640.

[27] 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 (NIPS)*, 2016, pp. 2234–2242.