

---

# HRVGAN: High Resolution Video Generation using Spatio-Temporal GAN

---

**Abhinav Sagar\***

Vellore Institute of Technology  
Vellore, Tamil Nadu, India  
abhinavsagar4@gmail.com

## Abstract

In this paper, we present a novel network for high resolution video generation. Our network uses ideas from Wasserstein GANs by enforcing k-Lipschitz constraint on the loss term and Conditional GANs using class labels for training and testing. We present Generator and Discriminator network layerwise details along with the combined network architecture, optimization details and algorithm used in this work. Our network uses a combination of two loss terms: mean square pixel loss and an adversarial loss. The datasets used for training and testing our network are UCF101, Golf and Aeroplane Datasets. Using Inception Score and Fréchet Inception Distance as the evaluation metrics, our network outperforms previous state of the art networks on unsupervised video generation.

## 1 Introduction

Deep learning for tackling computer vision problems has been mostly based on static image based approaches. However most real world data are dynamic in nature containing an additional time dimension which connects the images or individual frames together. Due to the presence of temporal dynamics, more information about the scene can be extracted. The challenge with video data is the additional computational burden and inherent complexity due to an additional time component. However, static image based algorithms are not suitable for action prediction problems (Huang et al., 2018). Hence, video based algorithms are the need for action prediction problems.

Neural networks for video generation from latent vectors is a challenging problem. State of the art methods produced blurry results thus showing the complexity of the problem (Vondrick et al., 2016). It is important to understand how pixels change in between the frames and model the uncertainty involved as shown in (Villegas et al., 2017). In case of video data, temporal dynamics needs to be separately modelled from the spatial dynamics. To infer various objects present in the scene spatial dynamics are used, where as the movement of these objects can be inferred from temporal dynamics. To solve this 1-D convolutions was used for temporal generator (Saito et al., 2017) and Recurrent Neural Networks (RNN) to generate latent code for image based generators (Tulyakov et al., 2018). Using 1-D convolutions reduces the computational burden, however for more accurate frame generation 3-D convolutions should be used.

However all of these previous work tackles very specific problems thus making generalization to other similar tasks difficult. Also almost all the architectures used in the literature, work for only specialized problems. Our work presents a novel unsupervised GAN based architecture for video generation/prediction which can be generalized to other settings.

---

\*Website of author - <https://abhinavsagar.github.io/>

## 2 Related Work

Generative models have been quite successful in modelling the time dynamics of video. Autoregressive models used in (Van den Oord et al., 2016) and Generative Adversarial Networks (GAN) used in (Xiong et al., 2018), (Acharya et al., 2018) and (Tulyakov et al., 2018) have enjoyed varying degrees of success. GANs have progressed in recent years due to better training stability as shown in (Salimans et al., 2016), better loss functions as used in (Deshpande et al., 2018) and improved architectures presented in (Karras et al., 2017). However work done using GANs for video data especially for action prediction problems is scarce.

Mode collapse is one of the fundamental challenges while training GANs in which the generated samples lack diversity ie the samples are similar to each other. Various approaches like using multiple generators used in (Ghosh et al., 2018) and using a reconstructor network which reverses the action of the generator by mapping from data to noise was used in (Srivastava et al., 2017). A new technique of progressively growing the network for stable generation of  $1024 \times 1024$  images was proposed in (Karras et al., 2017). Also a lot of work has been done to improve the individual loss functions of generator and discriminator as shown in (Arjovsky et al., 2017), (Deshpande et al., 2018) and (Gulrajani et al., 2017).

Video generation has been mostly tackled in a supervised setting in existing literature. Separating foreground from background was used for unsupervised video generation in (Vondrick et al., 2016). The architecture consists of two parallel streams consisting of 2D and 3D convolution layers for the generator and single stream 3D convolution layers for discriminator. A new approach using temporal and spatial generator was used in (Tulyakov et al., 2018). The videos generated using these methods were of resolution  $64 \times 64$ . A cascade architecture was used for unsupervised video generation in (Saito et al., 2017). Temporal generator consisting of 1-D deconvolution layers maps input latent vector to a set of new latent vectors corresponding to frames in the video. Each new latent vector is then fed to a new image generator for generating the video.

Our network differs from all of the previous approaches. We summarize our main contributions in this work as follows:

- We propose a GAN technique for unsupervised video generation of resolution  $256 \times 256$ .
- We present the architecture details of our network, optimization and loss functions used.
- We validate our network on publicly available UCF101 Dataset, Golf and Aeroplane Datasets for both qualitative and quantitative comparison.
- Our network beats the previous state of the art methods in this domain using Inception Score and Fréchet Inception Distance as the evaluation metrics.

## 3 Background

### 3.1 GAN

GANs are a family of unsupervised generative models which learns to generate samples from a given distribution (Goodfellow et al., 2014). Given a noise distribution, Generator  $G$  tries to generate samples while the Discriminator  $D$  tries to tell whether the generated samples are from the correct distribution or not. Both the generator and discriminator are trying to fool each other, thus playing a zero sum game. In other words both are in a state of Nash Equilibrium. Let  $G$  represent the generator and  $D$  the discriminator, loss function used for training GAN can be written as shown in Equation 1:

$$\mathcal{F}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [-\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [-\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))] \quad (1)$$

where  $z$  is latent vector,  $x$  is data sample,  $p_z$  is probability distribution over latent space and  $p_x$  is probability distribution over data samples. The zero sum condition is defined in Equation 2:

$$\min_G \max_D \mathcal{F}(\mathcal{D}, \mathcal{G}) \quad (2)$$

A lot of changes has been proposed over the years to reduce mode collapse and minimize training instability which are two of the main challenge while training GANs. Some of these changes are using least square loss instead of sigmoid cross entropy loss as shown in (Mao et al., 2017) and using feature matching and minibatch discrimination as shown in (Salimans et al., 2016).

### 3.2 Wasserstein GAN

A new technique was proposed to minimize Wasserstein Distance (WD) between the distributions to stabilize training. WD between two distributions was used in (Arjovsky et al., 2017) is defined in Equation 3:

$$W(p_r, p_g) = \inf_{\gamma \in \Pi(p_r, p_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (3)$$

where  $p_r, p_g$  are distributions of real and generated samples and  $\Pi(p_r, p_g)$  is the space of all possible joint probability distributions of  $p_r$  and  $p_g$ .

Another technique known as weight clipping was also proposed to enforce K-Lipschitz constraint. The loss function for training the network is defined as shown in Equation 4:

$$\mathcal{F}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\mathcal{D}(x)] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{u}}} [\mathcal{D}(\mathcal{G}(\mathbf{z}))] + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim p_{\mathbf{x}}} [(\|\nabla_{\hat{\mathbf{x}}} \mathcal{D}(\hat{\mathbf{x}})\|_2 - 1)^2] \quad (4)$$

Where  $\lambda$  is a regularization parameter.

### 3.3 Conditional GANs

These type of GANs use conditions on the generator in order to generate samples with desired property as first shown in (Mirza and Osindero, 2014). The loss functions for Conditional GANs can be defined as shown in Equation 5:

$$\mathcal{F}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [-\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{x}}} [-\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))] \quad (5)$$

The conditions could be class labels or original data sample in case of video prediction.

## 4 Method

### 4.1 Dataset

The following datasets were used in this work for training and testing our network for video generation:

1. UCF101 Dataset: The purpose of this dataset was training networks robust for action recognition tasks. It contains 13320 videos of 101 different action categories like Sky Diving, Knitting and Baseball Pitch (Soomro et al., 2012).
2. Golf and Aeroplane Datasets: It contains  $128 \times 128$  resolution frames which can be used for evaluating video generative adversarial networks (Vondrick et al., 2016) and (Kratzwald et al., 2017).

### 4.2 Network Architecture

Let input sequence frames of a video be denoted by  $(X = X_1, \dots, X_m)$  and frames to be predicted in sequence by  $(Y = Y_1, \dots, Y_n)$ . Our network for video generation has two stages: first: a new conditional generative adversarial network (GAN) to generate sequences performing a given category of actions, second: a reconstruction network with a new loss function to transfer sequences to the pixel space.

The input sequence frames in the form of noise vector is input to the Generator. The Generator generates output frames corresponding to the input frames. The output frame sequence is propagated to the Discriminator which tells whether the generated frames are real or fake. Both the Generator and

Discriminator is trained using mini batch Stochastic Gradient Descent(SGD) using the corresponding loss functions.

3D deconvolutional layers are used for generator and 3D convolutional layers for discriminator. Batch normalization is used for generator and instance normalization is used for discriminator network. ReLU activations is used as non linearity for generator and leaky ReLU activations for discriminator.

The network architecture used in this work is shown in Figure 1:

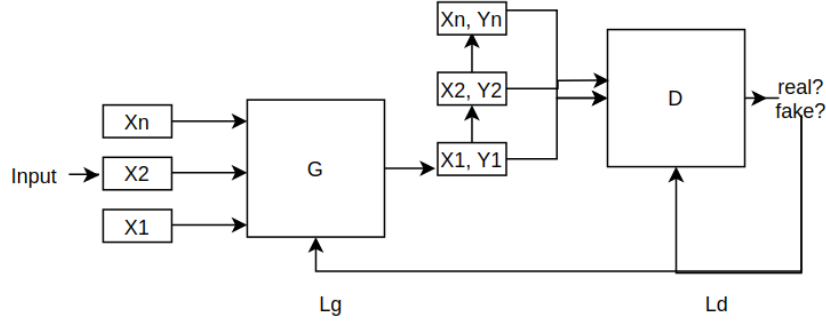


Figure 1: Network architecture used in this work

The Generator layer-wise details is shown in Table 1:

Table 1: Generator architecture for generation of  $256 \times 256 \times 32$  videos

Generator	Activation	Output shape
Latent vector	-	$128 \times 1 \times 1 \times 1$
Fully-connected	ReLU	$128 \times 1 \times 1 \times 1$
DeConv $3 \times 3 \times 3$	ReLU	$128 \times 4 \times 4 \times 4$
Upsample	-	$128 \times 8 \times 8 \times 8$
DeConv $3 \times 3 \times 3$	ReLU	$128 \times 8 \times 8 \times 8$
DeConv $3 \times 3 \times 3$	ReLU	$128 \times 8 \times 8 \times 8$
Upsample	-	$128 \times 8 \times 16 \times 16$
DeConv $3 \times 3 \times 3$	ReLU	$128 \times 8 \times 16 \times 16$
DeConv $3 \times 3 \times 3$	ReLU	$128 \times 8 \times 16 \times 16$
Upsample	-	$128 \times 8 \times 32 \times 32$
DeConv $3 \times 3 \times 3$	ReLU	$64 \times 8 \times 32 \times 32$
DeConv $3 \times 3 \times 3$	ReLU	$64 \times 8 \times 32 \times 32$
Upsample	-	$64 \times 16 \times 64 \times 64$
DeConv $3 \times 3 \times 3$	ReLU	$32 \times 16 \times 64 \times 64$
DeConv $3 \times 3 \times 3$	ReLU	$32 \times 16 \times 64 \times 64$
Upsample	-	$32 \times 16 \times 128 \times 128$
DeConv $3 \times 3 \times 3$	ReLU	$16 \times 16 \times 128 \times 128$
DeConv $3 \times 3 \times 3$	ReLU	$16 \times 16 \times 128 \times 128$
Upsample	-	$16 \times 32 \times 256 \times 256$
DeConv $3 \times 3 \times 3$	ReLU	$8 \times 32 \times 256 \times 256$
DeConv $3 \times 3 \times 3$	ReLU	$8 \times 32 \times 256 \times 256$
DeConv $1 \times 1 \times 1$	ReLU	$3 \times 32 \times 256 \times 256$

The Discriminator layer-wise details is shown in Table 2:

Table 2: Discriminator architecture for generation of  $256 \times 256 \times 32$  videos

Discriminator	Activation	Output shape
Input Image	-	$128 \times 1 \times 1$
Conv $1 \times 1 \times 1$	Leaky ReLU	$128 \times 4 \times 4 \times 4$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 4 \times 4 \times 4$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 4 \times 4 \times 4$
Downsample	-	$128 \times 8 \times 8 \times 8$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 8 \times 8 \times 8$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 8 \times 8 \times 8$
Downsample	-	$128 \times 8 \times 16 \times 16$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 8 \times 16 \times 16$
Conv $3 \times 3 \times 3$	Leaky ReLU	$128 \times 8 \times 16 \times 16$
Downsample	-	$128 \times 8 \times 32 \times 32$
Conv $3 \times 3 \times 3$	Leaky ReLU	$64 \times 8 \times 32 \times 32$
Conv $3 \times 3 \times 3$	LReLU	$64 \times 8 \times 32 \times 32$
Downsample	-	$64 \times 16 \times 64 \times 64$
Conv $3 \times 3 \times 3$	Leaky ReLU	$32 \times 16 \times 64 \times 64$
Conv $3 \times 3 \times 3$	Leaky ReLU	$32 \times 16 \times 64 \times 64$
Downsample	-	$32 \times 16 \times 128 \times 128$
Conv $3 \times 3 \times 3$	Leaky ReLU	$16 \times 16 \times 128 \times 128$
Conv $3 \times 3 \times 3$	Leaky ReLU	$16 \times 16 \times 128 \times 128$
Downsample	-	$16 \times 32 \times 256 \times 256$
Minibatch Stddev	-	$129 \times 4 \times 4 \times 4$
Conv $3 \times 3 \times 3$	Leaky ReLU	$8 \times 32 \times 256 \times 256$
Fully-connected	linear	$1 \times 1 \times 1 \times 128$
Fully-connected	linear	$1 \times 1 \times 1 \times 1$

### 4.3 Pixel Normalization

To avoid explosion of parameters in both generator and discriminator, feature vectors are normalized at every pixel. We extended the feature vector normalization as proposed by (Karras et al., 2017) to our spatio-temporal problem.

Let  $a_{x,y,t}$  and  $b_{x,y,t}$  be original and normalized feature vector at pixel  $(x, y, t)$  corresponding to spatial and temporal position. The following relation can be written as shown in Equation 6:

$$b_{x,y,t} = \frac{a_{x,y,t}}{\sqrt{\frac{1}{N} \sum_{j=0}^{N-1} (a_{x,y,t}^j)^2 + \epsilon}} \quad (6)$$

where  $\epsilon$  is a constant and  $N$  is number of feature maps used.

### 4.4 Instance Normalization

Instance normalization was used after both 3D convolutional and 3D deconvolutional layers to solve the vanishing gradient problem as defined in Equation 7.

$$y = \text{ReLU} \left( \sum_{i=0}^d w_i \cdot \text{ReLU} \left( \gamma_i \cdot \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i \right) + b \right) \quad (7)$$

where  $w$  and  $b$  are weight and bias term of the 3D convolution layer,  $\gamma$  and  $\beta$  are weight and bias term of the Instance Normalization layer,  $\mu$  and  $\sigma$  are mean and variance of the input.

#### 4.5 Loss Functions

Generator in the GAN architecture can be used to predict sequence of frames  $Y$  from sequence of frames  $X$  by minimizing the pixel wise distance between the predicted and the actual frame. The mean square pixel wise loss function is defined in Equation 8:

$$\mathcal{L}_{mse}(X, Y) = \ell_{mse}(G(X), Y) = \|G(X) - Y\|^2 \quad (8)$$

The binary cross-entropy loss between the actual and predicted frames is defined in Equation 9:

$$L_{bce}(Y, \hat{Y}) = - \sum_i \hat{Y}_i \log(Y_i) + (1 - \hat{Y}_i) \log(1 - Y_i) \quad (9)$$

where both  $Y_i$  and  $\hat{Y}_i$  has values in the range  $[0, 1]$ .

Let  $(X, Y)$  be a sample from the dataset where both  $X$  and  $Y$  denote a sequence of frames as input and to be predicted respectively. Let  $G$  represent the Generator and  $D$  the Discriminator. The goal is to predict the right frames for both the individual classes represented by 0 and 1. The adversarial loss function used for training Generator is defined in Equation 10:

$$\mathcal{L}_{adv}^G(X, Y) = \lambda_1 \sum_{i=1}^N L_{bce}(D_i(X_i, G_i(X_i)) - k, 1) \quad (10)$$

The adversarial loss function used for training the Discriminator is defined in Equation 11:

$$\mathcal{L}_{adv}^D(X, Y) = \lambda_1 \sum_{i=1}^N L_{bce}(D_i(X_i, Y_i) - k, 1) + \lambda_2 \sum_{i=1}^N L_{bce}(D_i(X_i, G_i(X_i)) - k, 0) \quad (11)$$

Where  $\lambda_1, \lambda_2$  are the coefficients to balance the penalty terms.  $\lambda_1, \lambda_2$  are also used to absorb the scale  $k$  caused by the k-Lipschitz constraint on Wasserstein loss.

The mean square loss function in Equation 6 and adversarial loss function in Equation 8 of Generator can be combined with equal weights given to both the terms as shown in Equation 12:

$$\mathcal{L}(X, Y) = \alpha \mathcal{L}_{adv}^G(X, Y) + \beta \mathcal{L}_{mse}(X, Y) \quad (12)$$

Where  $\alpha$  and  $\beta$  are constants.

#### 4.6 Algorithm

The complete algorithm used in this work is shown below:

---

**Algorithm 1:** HRVGAN: High Resolution Video Generation using Spatio-Temporal GAN

---

Initialize learning rates  $\alpha_D$  and  $\alpha_G$ , and weights  $\lambda_{adv}, \lambda_{mse}$

**while not converged do**

**Update the Discriminator  $D$ :**

    Get  $M$  data samples  $(X, Y) = (X^{(1)}, Y^{(1)}), \dots, (X^{(M)}, Y^{(M)})$

$W_D = W_D - \alpha_D \sum_{i=1}^M \frac{\partial \mathcal{L}_{adv}^D(X^{(i)}, Y^{(i)})}{\partial W_D}$

**Update the Generator  $G$ :**

    Get  $M$  data samples  $(X, Y) = (X^{(1)}, Y^{(1)}), \dots, (X^{(M)}, Y^{(M)})$

$W_G = W_G - \alpha_G \sum_{i=1}^M \left( \lambda_{adv} \frac{\partial \mathcal{L}_{adv}^G(X^{(i)}, Y^{(i)})}{\partial W_G} + \lambda_{mse} \frac{\partial \mathcal{L}_{mse}(X^{(i)}, Y^{(i)})}{\partial W_G} \right)$

**end**

---

## 4.7 Evaluation Metrics

A lot of metrics have been proposed for evaluating GANs in the literature. Two of the most common metrics are Inception Score and Fréchet Inception Distance which are explained below:

1. **Inception Score (IS)** - Inception Score was first proposed in (Salimans et al., 2016) for evaluating GANs. A higher inception score is preferred which means the model is able to generate diverse images thus avoiding mode collapse issue.

Let  $x$  be samples generated by the generator  $G$ ,  $p(y|x)$  be the distribution of classes for generated samples and  $p(y)$  be the marginal class distribution. The Inception score is defined as in Equation 13:

$$IS(\mathcal{G}) = \exp \left( \mathbb{E}_{\mathbf{x} \sim p_g} \mathcal{D}_{KL}(p(y | \mathbf{x}) || p(y)) \right) \quad (13)$$

where  $D_{KL}$  is the Kullback-Leibler divergence between  $p(y|x)$  and  $p(y)$ .

2. **Fréchet Inception Distance (FID)** - Another metric to evaluate the quality of generated samples was first proposed by (Heusel et al., 2017).

Let  $D$  represent the CNN used to extract features,  $(m_r, \sigma_r)$  be mean and covariance of features extracted from real samples and  $(m_f, \sigma_f)$  be mean and covariance of features extracted from fake samples with  $D$ , then the Fréchet Inception distance is defined as in Equation 14:

$$d^2((m_r, \Sigma_r), (m_f, \Sigma_f)) = \|m_r - m_f\|_2^2 + \text{Tr} \left( \Sigma_r + \Sigma_f - 2(\Sigma_r \Sigma_f)^{1/2} \right) \quad (14)$$

Fréchet Inception Distance is more accurate than Inception Score as it compares summary statistics of generated samples and real samples. A lower FID is preferred for better performing generative models.

## 5 Results

The Inception scores of ours compared with other models on UCF101 dataset is shown in Table 3:

Table 3: Inception scores of ours compared with other models on UCF101 dataset

Model	Inception Scores
VGAN (Vondrick et al., 2016)	8.18
TGAN (Saito et al., 2017)	11.85
MoCoGAN (Tulyakov et al., 2018)	12.42
Ours	14.29

We next present the quantitative comparison of our network with TGAN and VideoGAN based on FID score on Golf and Aeroplane datasets in Table 4:

Table 4: Quantitative comparison of Our network with TGAN and VideoGAN based on FID score on Golf and Aeroplane datasets

Model	FID Score on Golf Dataset	FID Score on Aeroplane Dataset
VGAN (Vondrick et al., 2016)	113007	149094
TGAN (Saito et al., 2017)	112029	120417
Ours	102584	104036

The linear interpolation in latent space to generate samples from Golf dataset is shown in Figure 2:

The linear interpolation in latent space to generate samples from Aeroplane dataset is shown in Figure 3:

The generated frames using UCF-101 dataset is shown in Figure 4:



Figure 2: Linear interpolation in latent space to generate samples from Golf dataset

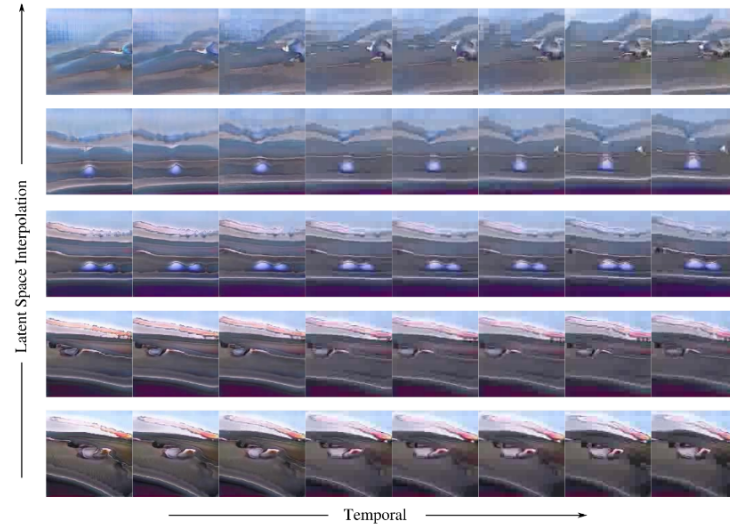


Figure 3: Linear interpolation in latent space to generate samples from Aeroplane dataset



Figure 4: Results on UCF-101 generated from random noise. For each task, we display 8 frames of our generated videos for the JumpingJack (1st row) and TaiChi (2nd row).



## 6 Conclusions

In this paper, we presented a novel neural network using generative models for unsupervised video generation. Our network is an extension of original GAN architecture which is trained using mini batch Stochastic Gradient Descent. The novel loss term is made up of a mean square pixel loss along with an adversarial loss which uses k-Lipschitz constraint on it as used in Wasserstein GANs. We present the architecture details, optimization and the complete algorithm used in this work. On testing our network on UCF101, Golf and Aeroplane Datasets using Inception Score and Fréchet Inception Distance as the evaluation metrics, our network outperforms previous state of the art approaches. Finally we also present the linear interpolation in latent space on Golf and Aeroplane Datasets and the frames generated using UCF101 dataset.

## Acknowledgments

We would like to thank Nvidia for providing the GPUs for this work.

## References

- D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv preprint arXiv:1810.02419*, 2018.
- M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018.
- J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- A. Clark, J. Donahue, and K. Simonyan. Efficient video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.
- E. Denton and R. Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.
- I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491, 2018.
- A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. Torr, and P. K. Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8513–8521, 2018.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.
- J. He, A. Lehrmann, J. Marino, G. Mori, and L. Sigal. Probabilistic video generation using holistic attribute control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.
- D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018.

- T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- B. Kratzwald, Z. Huang, D. P. Paudel, and L. Van Gool. Towards an understanding of our world by ganing videos in the wild. *CoRR, abs/1711.11453*, 2017.
- X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.
- L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.
- M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- J. Pan, C. Wang, X. Jia, J. Shao, L. Sheng, J. Yan, and X. Wang. Video generation from single semantic label map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2019.
- A. Sagar. Bayesian multi scale neural network for crowd counting. *arXiv preprint arXiv:2007.14245*, 2020.
- M. Saito and S. Saito. Tganv2: Efficient training of large models for video generation with multiple subsampling layers. *arXiv preprint arXiv:1811.09245*, 2018.
- M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017.
- T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.
- K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.
- X. Sun, H. Xu, and K. Saenko. A two-stream variational adversarial network for video generation. *arXiv preprint arXiv:1812.01037*, 2018.
- R. R. Torrado, A. Khalifa, M. C. Green, N. Justesen, S. Risi, and J. Togelius. Bootstrapping conditional gans for video game level generation. *arXiv preprint arXiv:1910.01603*, 2019.
- S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.
- A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.
- R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831*, 2017.
- C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.
- C. Wang, C. Xu, C. Wang, and D. Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27(8):4066–4079, 2018.

- Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020.
- W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.
- C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.