ImaGINator: Conditional Spatio-Temporal GAN for Video Generation

Yaohui Wang, Piotr Bilinski, Francois F Bremond, Antitza Dantcheva

▶ **To cite this version:**

Yaohui Wang, Piotr Bilinski, Francois F Bremond, Antitza Dantcheva. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. WACV 2020 - Winter Conference on Applications of Computer Vision, Mar 2020, Snowmass Village, United States. hal-02368319

HAL Id: hal-02368319
https://hal.science/hal-02368319v1

Submitted on 18 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ImaGINator: Conditional Spatio-Temporal GAN for Video Generation

Yaohui Wang[1,2] Piotr Bilinski[3] Francois Bremond[1,2] Antitza Dantcheva[1,2]
[1]Inria [2] Université Côte d'Azur [3]University of Warsaw
{yaohui.wang, francois.bremond, antitza.dantcheva}@inria.fr bilinski@mimuw.edu.pl

Abstract

Generating human videos based on single images entails the challenging simultaneous generation of realistic and visual appealing appearance and motion. In this context, we propose a novel conditional GAN architecture, namely ImaGINator, which given a single image, a condition (label of a facial expression or action) and noise, decomposes appearance and motion in both latent and high level feature spaces, generating realistic videos. This is achieved by (i) a novel spatio-temporal fusion scheme, which generates dynamic motion, while retaining appearance throughout the full video sequence by transmitting appearance (originating from the single image) through all layers of the network. In addition, we propose (ii) a novel transposed (1+2)D convolution, factorizing the transposed 3D convolutional filters into separate transposed temporal and spatial components, which yields significantly gains in video quality and speed. We extensively evaluate our approach on the facial expression datasets MUG and UvA-NEMO, as well as on the action datasets NATOPS and Weizmann. We show that our approach achieves significantly better quantitative and qualitative results than the state-of-the-art. The source code and models are available under https://github.com/wyhsirius/ImaGINator.

1. Introduction

Generating realistic human videos based on single images brings to the fore following three challenges: (a) retaining of human appearance throughout the video, (b) generating (uncertain) motion, as well as (c) modeling of spatio-temporal consistency. Finding suitable representation learning methods, which are able to address these challenges, is critical to the final visual quality and plausibility of the rendered novel video sequences.

Existing methods predominantly treat generation of high dimensional video as a separate *two step* modeling of low-dimensional temporal and spatial generation. Such methods (*e.g.* MoCoGAN) [37], are grounded on the *seq2seq* [35] architecture. In particular associated video generation in such methods includes two steps: (1) motion generation in a latent space, proceeded by (2) motion and appearance-generation, where frames are generated individually, combining the single-input-image-appearance information with each motion vector generated in (1). These two steps aim at decomposing video generation into the generation of individual frames, which imparts the benefit of straightforward optimization. Two step methods fail to address the above named challenges (a) and (c), *i.e.* appearance is not sufficiently retained and spatio-temporal consistency is not modeled, as temporal consistency is not modeled in higher level spatial spaces.

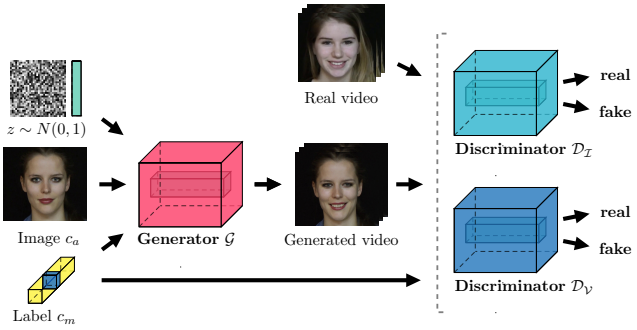


Figure 1: The proposed ImaGINator architecture incorporates *Generator* G, *image Discriminator* D_I, as well as *video Discriminator* D_V. G accepts c_a, c_m and noise as input, and seeks to generate realistic video sequences. While D_I discriminates whether the generated images contain an authentic appearance, D_V additionally determines whether the generated videos contain an authentic motion.

In contrast to two step methods, VGAN [39] utilized a *single step* to generate future frames by leveraging on 3D convolution to model spatio-temporal features in high and low levels. We here note that utilizing 3D convolution directly challenges optimization. In addition, the generated video was decomposed into foreground and background, in two streams, which required an additional branch to model background information, increasing the complexity of the model.

Motivated by the above, we propose a new conditional GAN model, referred to as ImaGINator, generating video sequences given a single image c_a, a motion class c_m (*i.e.* facial expression or human action), as well as noise. ImaGINator incorporates a *Generator* G, a *video Discriminator*

D_V, as well as an *image Discriminator* D_I, as depicted in Figure 1. While the Generator G, based on a fully convolutional 'Encoder-Decoder'-architecture, accepts c_a and c_m as input to generates video sequences, the image Discriminator D_I and the video Discriminator D_V assess the authenticity of appearance and motion of generated videos.

ImaGINator is streamlined to exploit the joint benefits of single and two-step methods by incorporating following new properties.

- A *novel spatio-temporal fusion* mechanism, aiming at *retaining the appearance* by enforcing G to employ the spatial information in both, low and high feature levels. By injecting c_a into the *Decoder*, we enable G to place emphasis on generating solely motion. This is based on the hypothesis that a video can be disentangled into appearance and motion in the latent space, as well as in multi-level spatio-temporal feature spaces. While at each level appearance is retained, only the motion is being altered.
- A *novel transposed (1+2)D convolution*, factorizing the transposed 3D convolutional filters into separate temporal and spatial components. This brings several benefits: (i) an additional nonlinear rectification allows the model to represent more complex functions, (ii) it facilitates optimization, as transposed (1+2)D convolution blocks are easier to optimize than the full transposed 3D convolutional filters, and (iii) it yields significant gains in both video quality and speed.

Towards comparing our algorithm with other video generation algorithms, we augment two state-of-the-art video generation algorithms, namely VGAN and MoCoGAN, in order to adhere to our image-to-video-generation-setting. We proceed to provide a comparison, showing that our method outperforms these methods *qualitatively* (based on a human study of 30 subjects) and *quantitatively* on both, facial expression (MUG and UvA-NEMO), as well as human action datasets (Weizmann and NATOPS) by presenting results pertaining to five evaluation metrics. In addition, we conduct an ablation study, which validates the effectiveness of components in ImaGINator.

We note that while ImaGINator can be generally applied to many domains of computer vision, we here present experiments in the language of facial expression and action generation. Specifically, we focus on the setting, where we provide jointly a single frame of a subject, defining the appearance, and a condition (*i.e*. label), determining facial expression or human action, and proceed to generate a video exhibiting the subject from the initial frame performing the named expression or action.

2. Related Work

Conditional Generation accepts as inputs both, latent variables, as well as known auxiliary information, such as class labels. The majority of works have expanded either Generative Adversarial Networks (GANs) [9] or Variational Auto-Encoders (VAEs) [18] in this context, by augmenting GANs and VAEs with the capability of generating data samples based on class labels. Conditional generation has been beneficial in domain transfer, super-resolution imaging, video to video translation, as well as image and face editing [13, 54, 26, 15, 20, 43, 4, 16, 45, 46, 50, 44]. Most recently, a number of new techniques have been proposed to stabilize the training process of conditional GANs (cGANs) and improve the visual quality of generated images [27, 3]. Our proposed ImaGINator is a cGAN architecture, aiming at generating facial expressions / human actions from single images, where a category label is provided in both G and D.

Unsupervised **video prediction based on multiple frames** involves the use of multiple frames as input and the prediction of future frames by learning to extrapolate. Video prediction has been predominantly focused on predicting high-level semantics in video, such as action [33, 19, 8, 25, 38, 47, 6, 5], event [51, 12, 32], semantic segmentation [24], as well as motion [30, 41, 40, 22]. In contrast to such works, our model is targeted to generate a video sequence based on a *single frame*. Since future motion is very uncertain under this setting, we leverage action label as a guidance.

Video generation based on a single image is challenging and hence current methods have proposed to decompose it into sub-tasks. One line of scientific works have utilized in this additional context-information, *e.g*. human key points [14, 49, 42], 3D face mesh [52] and optical flow [21], as future motion guidance. This additional information is either pre-computed throughout the generated video [14, 52] or predicted based on an initial input [49, 42]. The additional information guides a conditional image translation, which though results in lack of modeling spatio-temporal correlations.

Deviating from the above, MoCoGAN [37], VGAN [39] and Xue *et al*. [48] attempted to hallucinate future frames directly in the pixel space. The latter proposed a probabilistic model, predicting dynamic filters on the input image to render next frame, leading to prediction of only one future frame. MoCoGAN is based on a *seq2seq* [35] architecture, aiming at separating spatio-temporal generation into two steps (disentangling each video frame into motion and appearance in different latent spaces). However, such two-step generation omits the modeling of temporal consistency in higher spatial levels, which generally fails to retain original appearance. VGAN employs a single step method towards modeling multi-level spatio-temporal consistency through 3D convolution by decomposing videos into foreground and background. Although it models both, low and high level features, due to lack of frame quality constrains, generated results are of inherently lower visual quality, *i.e*. are blurry.

Deviating from the above, we propose a single step ar-

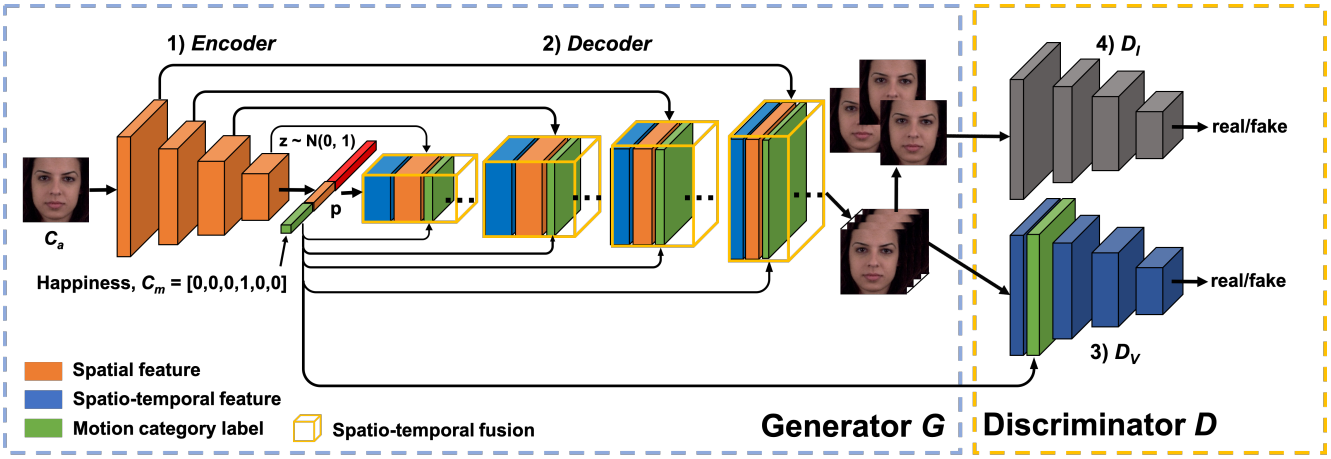


Figure 2: **Overview of the proposed ImaGINator**. In the *Generator* G, the *Encoder* firstly encodes an input image c_a into a single vector p. Then, the *Decoder* produces a video based on a motion c_m and a random vector z. By using spatio-temporal fusion, low level spatial feature maps from the *Encoder* are directly concatenated into the *Decoder*. While D_I discriminates whether the generated images contain an authentic appearance, D_V additionally determines whether the generated videos contain an authentic motion.

chitecture, which decomposes motion and appearance in multi-level feature spaces for image to video generation.

The rest of the paper is organized as follows. In Section 3 we introduce the new ImaGINator framework. Qualitative and quantitative analyses of our model are presented in Section 4. Section 5 concludes the paper and provides directions for future research.

3. Proposed Approach

Our goal is to generate a video sequence given an appearance information (as a single image frame) and a motion class (*e.g.* determining the facial expression). We here assume that a video y can be decomposed into appearance c_a (originating from the input-image) and motion c_m (originating from the category-label), based on which we proceed to generate videos. Hence, we formulate our task as learning a conditional mapping $G\colon \{z, c_a, c_m\} \rightarrow y$, where $z \sim \mathcal{N}(0,\, 1)$ denotes the random noise.

Towards achieving our goal, we propose a framework that consists of the following 3 main components: (i) *Generator* G, that accepts c_a, c_m and noise as inputs, and seeks to generate realistic video sequences, (ii) *image Discriminator* D_I that determines the frame-level based appearance quality, and (iii) *video Discriminator* D_V, which additionally discriminates, whether the generated video sequences contain authentic motion, see Figure 1.

3.1. Network Architecture

In the following we proceed to describe the architecture of our video prediction network, providing details on G, D_I and D_V, as illustrated in Figure 2. In addition, we elaborate on the proposed spatio-temporal fusion scheme, as well as the transposed (1+2)D convolution.

3.1.1 Generator

Our Generator G consists of an image *Encoder* and a video *Decoder*, see Figure 2. The *Encoder* extracts appearance information in various layers, from shallow, fine layers to deep, coarse layers. It encodes the input image c_a into a latent vector p, and then by concatenating p, c_m as well as the random noise $z \sim \mathcal{N}(0, 1)$, the decoder generates a video sequence.

In our Generator G, we extend the idea of using 2 skip connections from the FCN-8 [23] to 4 skip connections, but with the difference that the original skip connections are applied to fuse predictions, whereas ours are applied to fuse appearance and motion spatio-temporal features. Our skip connections allow the Decoder to access low-level features directly from the Encoder, enabling the Decoder to reuse the appearance features at each time slice and to focus on generating motion.

Spatio-temporal fusion. Let G have n layers and let $F_i^{H\times W\times C_1\times T}$ be the feature map from the i^{th} layer with C_1 number of channels in G, $f_{i,t}^{H\times W\times C_1}, t \in \{1, ..., T\}$ be the t^{th} feature map in F_i and $F_{n-i}^{H\times W\times C_2}$ represent the feature map from $(n-i)^{th}$ layer, see Figure 3. We design the outputs of each respective layer from our Decoder and Encoder to have the same spatial dimensions $H \times W$. We propose a fusion mechanism, concatenating each $f_{i,t}$ with F_{n-i} in a *channel-wise dimension* with a result of a new feature map $F_i^{'H\times W\times (C_1+C_2)\times T}$, named **spatio-temporal fusion**. Here we note that each initial feature map F_i represents spatio-temporal features of several consecutive frames in the generated video. By spatio-temporally fusing F_i and F_{n-i} directly in *different feature levels*, the input information can be well preserved in the generated video.

Further, we fuse the category label (constituting a one-

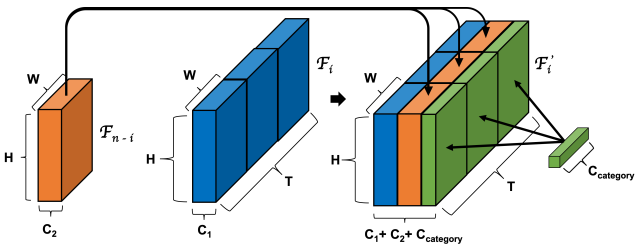


Figure 3: **Spatio-temporal fusion.** Blue and orange cuboids represent the intermediate feature maps in the Decoder and Encoder respectively. Our proposed fusion scheme enforces the Decoder reutilizing spatial information through skip connections. Based on such operations, temporal consistency can be modeled in multi-levels.

hot vector) directly into the Decoder, in order to provide each layer an access to the label. To do so, we firstly project the one-hot vector onto one-hot feature map. Then, we spatio-temporally fuse the category label information into different layers in the Decoder. Our final feature map F_i is of size $H \times W \times (C_1 + C_2 + C_{category}) \times T$.

We note that 3D convolution, utilized in one step methods often brings to the fore generation of blurry videos, due to hard optimization. Nevertheless, benefiting from spatial and temporal decomposition, frames can be generated individually in a two step method. Hence, towards incorporating such decomposition in a one step method, we design a new convolution layer, integrating transposed (1+2)D convolution.

Transposed (1+2)D Convolution. We propose to explicitly factorize transposed 3D convolutional filters into two separate and successive operations, M transposed 1D temporal convolutional filters followed by a 2D separate spatial components, which we refer to as transposed (1+2)D convolution, shown in Figure 4. Such decomposition brings to the fore several benefits. The first benefit relates to an additional nonlinear rectification between these two operations, thus allowing the model to represent more complex functions. The second potential benefit is that the decomposition facilitates optimization, as transposed (1+2)D convolution blocks, with factorized temporal and spatial components, are easier to optimize than the full transposed 3D convolutional filters. Moreover, we show that factorizing the transposed 3D convolutional filters yields significant gains in both, video quality and speed, see Section 4. We note that proposed transposed (1+2)D convolution is inspired by decomposition of 3D convolutional filters [36].

3.1.2 Two-stream Discriminator

Towards improving image quality in video generation, we here design a two-stream *Discriminator* architecture, containing D_V, as well as D_I. While D_V has five 3D convolution layers, D_I contains only 2D convolutions with the same

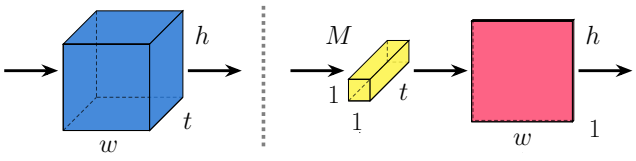


Figure 4: **Transposed 3D convolution** (on the left) *vs*. **proposed Transposed (1+2)D convolution** (on the right). The transposed 3D convolutional filter of size $t \times w \times h$ has been decomposed into M transposed 1D temporal convolution filters $t \times 1 \times 1$ and a transposed 2D spatial convolution $1 \times w \times h$. The operation M denotes the number of 1D filters, t indicates the temporal size, and w and h indicate the spatial size.

layer numbers of D_V. D_V accepts the full generated video as input, using proposed spatio-temporal fusion to fuse the 'one-hot feature map' of the category label and the output of the first layer, similarly like in G. D_V seeks to measure the KL divergence between the joint distributions $p(x_{real}, c_m)$ and $p(x_{fake}, c_m)$. We randomly sample N frames out of real and generated video respectively as input.

3.2. Formulation

Our goal is to learn the mapping function G, *i.e.* $G\colon \{z, c_a, c_m\} \rightarrow y$, given training samples. In addition, we introduce two adversarial discriminators D_I and D_V.

Full Objective. We define our full objective function as

$$\mathcal{L}_{\mathcal{F}}(G, D_I, D_V) = \mathcal{L}_{GAN}(G, D_I, D_V) + \lambda \mathcal{L}_{rec}(G), \quad (1)$$

which contains two types of terms: an *adversarial loss* $\mathcal{L}_{GAN}$ for matching the distribution of generated images to the data distribution in the target domain, and a *reconstruction loss* $\mathcal{L}_{rec}$ for capturing the overall structure and coherence of a video. Due to the high dimensional video space, we introduce the λ parameter, which controls the relative importance of the objectives and stabilizes the training and balancing between losses. We aim to solve

$$G^* = \arg\min_{G} \max_{D_I, D_V} \mathcal{L}(G, D_I, D_V). \quad (2)$$

Adversarial Losses. We apply adversarial losses [9] to our mapping function G and its image Discriminator D_I and video Discriminator D_V. We express the objective as

$$\mathcal{L}_{GAN}(G, D_I, D_V) = \mathcal{L}_I(G, D_I) + \mathcal{L}_V(G, D_V), \quad (3)$$

where G attempts to generate videos $G(z, c_a, c_m)$, which resemble real videos from domain Y, while D_I and D_V aim to distinguish between translated samples $G(z, c_a, c_m)$ and real samples $y \in Y$. G seeks to minimize this objective against adversaries D_I and D_V, which attempt to maximize it, *i.e.* $\min_G \max_{D_I, D_V} \mathcal{L}_{GAN}(G, D_I, D_V)$. The loss $\mathcal{L}_I$

and the loss $\mathcal{L}_V$ are defined as follows.

$$\mathcal{L}_I = \mathbb{E}_{x' \sim p_{data}}[log(D_I(x^{'}))] + \mathbb{E}_{z \sim p_z(z), c_a, c_m}[1 - log(D_I(G(z, c_a, c_m)^{'}))], \quad (4)$$

$$\mathcal{L}_V = \mathbb{E}_{x \sim p_{data}, c_m}[log(D_V(x, c_m))] + \mathbb{E}_{z \sim p_z(z), c_a, c_m}[1 - log(D_V(G(z, c_a, c_m), c_m))]. \quad (5)$$

$\mathcal{L}_I$ denotes the loss function related to D_I, $\mathcal{L}_V$ represents the loss function related to D_V, and $(\cdot)^{'}$ characterizes N frames sampled from real and generated videos. Both losses, encompassed in D_I and D_V, are based on the *Cross-Entropy* loss.

Reconstruction Loss. We define our video-level reconstruction loss as

$$\mathcal{L}_{rec} = \mathbb{E}[||x_{real} - G(z, c_a, c_m)||_1]. \quad (6)$$

The reconstruction loss is aimed at capturing the overall structure and coherence of a video. It uses $\mathcal{L}_1$ loss in order to generate sharp videos. By combining it with $\mathcal{L}_{GAN}$, it fosters G to create more realistic videos and to reconstruct the original real ones at the same time.

Ablation study. In the supplementary material, we compare our method against ablations of the full objective, including the adversarial loss $\mathcal{L}_{GAN}$ alone and the video-level reconstruction loss $\mathcal{L}_{rec}$, empirically showing that both objectives play critical roles in contributing to obtained accuracy.

Training strategy. To train the network, we firstly provide an input frame, as well as corresponding category label to G to generate possible videos. Then D_V and D_I distinguish between real and fake videos and frames based on the respective quality. Specifically, when training D_V, we provide two types of negative samples, generated videos with correct labels $(x_{real}, c_{correct})$ and real videos with wrong labels (x_{real}, c_{wrong}). We observe that such training enforces D_V to learn from diverse samples and at the same time enables the generation of realistic samples. We provide details in Algorithm 1.

Algorithm 1 ImaGINator Training Algorithm

Input: minibatch x, $x^{'}$, input image c_a, correct c_m, wrong $\hat{c_m}$

for each step **do**
- $z \sim \mathcal{N}(0, I)$
- $x_{recon} \leftarrow G(z, c_a, c_m)$
- $s_{real} \leftarrow D_V(x, c_m) + D_I(x^{'})$
- $s_{recon} \leftarrow D_V(x_{recon}, c_m) + D_I(x^{'}_{recon})$
- $s_w \leftarrow D_V(x, \hat{c_m}) + D_I(x^{'})$
- $\mathcal{L}_D \leftarrow \log(s_r) + 0.5[\log(1 - s_w) + \log(1 - s_{recon})]$
- $D_V \leftarrow D_V - \alpha \partial \mathcal{L}_D / \partial D_V$
- $D_I \leftarrow D_I - \alpha \partial \mathcal{L}_D / \partial D_I$
- $\mathcal{L}_{recon} \leftarrow ||x - x_{recon}||_1$
- $\mathcal{L}_G \leftarrow \log(s_{recon}) + \lambda \mathcal{L}_{recon}$
- $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$

end for

4. Experiments

Experimental Setup. We train the entire network end-to-end with the standard back-propagation algorithm using only a single NVIDIA GeForce GTX 1080Ti with 11 GB of memory. We employ ADAM optimizer [17] with $\beta = 0.5$. Moreover, we apply spectral normalization on both D_I and D_V to stabilize training, as proposed by Miyoto *et al*. [27]. We observe that given the same learning rate for D_I, D_V and G during training, D_I and D_V typically learn faster than G. The reason for this might be that the spatio-temporal convolution is more efficient at differentiating than at generating, as pointed out by Goodfellow *et al*. [9] and Radford *et al*. [31]. In order to circumvent this disparity, we set the learning rate to $2e^{-4}$ for G, and $5e^{-5}$ for both D_I and D_V. λ is set $1e^{-3}$ to balance two types of losses.

4.1. Datasets

We comprehensively evaluate our method on the following four datasets.

The **MUG** Facial Expression dataset [1] contains 931 videos of 52 subjects (data of 42 subjects is employed for training and 10 for testing), performing 7 facial expressions, namely "happy", "sad", "surprise", "anger", "disgust", "fear" and "neutral".

The **NATOPS** Aircraft Handling Signals dataset [34] contains video sequences of 20 subjects (data of 15 subjects is employed for training and 5 for testing), performing 24 gestures including "all clear" and "move ahead". Each subject repeats each gesture 20 times.

The **Weizmann** Action dataset [10] contains 90 videos of 9 subjects (data of 6 subjects is employed for training and 3 for testing), performing 10 actions, *e.g*. "wave" and "bend". We augment this dataset by doubling the number of videos using horizontal flipping transformation.

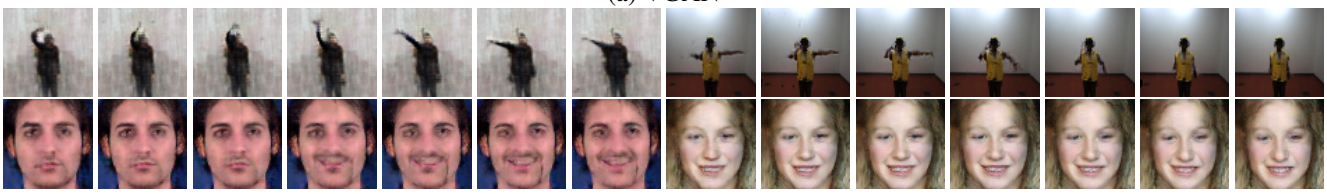
The **UvA-NEMO** Smile dataset [7] contains 597 video sequences of smiling individuals. It contains 400 subjects (data of 320 subjects is employed training and 80 for testing) with 1 or 2 videos per subject. In the context of UvA-NEMO we do not provide any category to our model, since the dataset features only one facial expression.

In all our experiments, images are scaled to 64×64 pixels. We use a time step 2 to sample frames from facial expression datasets and a time step of 1 from human action datasets. MUG and UvA-NEMO are pre-processed by detecting faces in OpenFace [2] and cropping them in each frame.

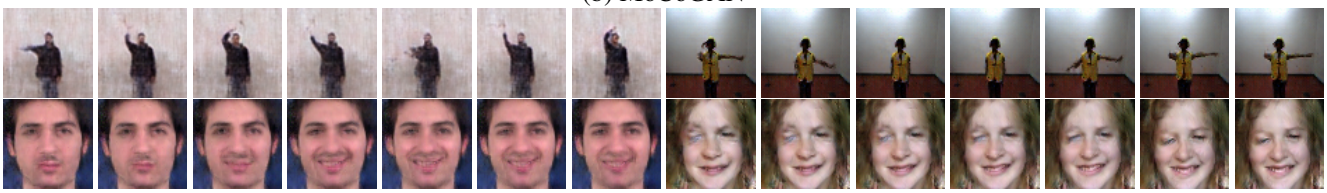
4.2. Evaluation Metrics

The Video Fréchet Inception Distance (**FID**) [43] is a video generation metric. It measures both visual quality and temporal consistency of generated videos. We use 3D ResNeXt-101 [11] as a feature extractor and calculate Video

(a) VGAN



(b) MoCoGAN



(c) Proposed ImaGINator

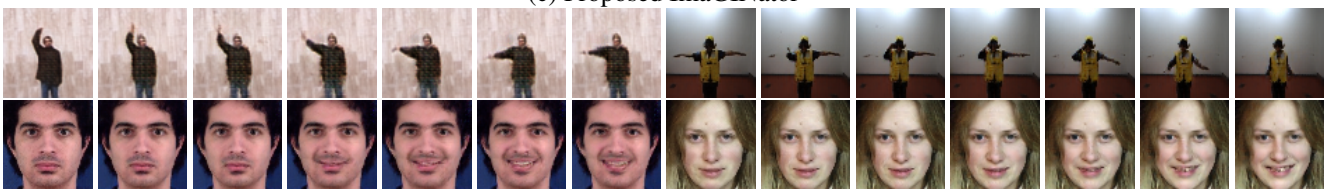


Figure 5: **Example generated video frames** pertained to algorithms (a) VGAN, (b) MoCoGAN, as well as the (c) proposed ImaGINator. For each method, we present generated video frames for the four datasets: **Weizmann** (top-left), label *"Wave"*; **NATOPS** (top-right), label *"Hot Brakes"*; **MUG** (bottom-left), label *"Happiness"*; **UvA-NEMO** (Down-right), no label. All frames are sampled with a time step of 3.

FID as: $\|\mu - \widetilde{\mu}\|^2 + Tr(\Sigma + \widetilde{\Sigma} - 2\sqrt{\Sigma\widetilde{\Sigma}}))$, where μ and Σ are mean and covariance matrix computed from real feature vectors, and $\widetilde{\mu}$, and $\widetilde{\Sigma}$ are computed from generated data. Lower Video FID scores represent a superior quality of generated videos.

The Structural Similarity Index Measure (**SSIM**) indicates the structure similarity between real and reconstruction images, Peak Signal-to-Noise Ratio (**PSNR**) quantifies the image quality. High SSIM and PSNR scores indicate higher quality of generated images.

The Average Content Distance (**ACD-C**) [37] measures content consistency of a generated video. For facial expression videos, we first use OpenFace [2], which outperforms human performance in face recognition, to extract a feature vector pertaining to the detected face. Then, we compute the ACD-C as an average $\mathcal{L}_2$ pairwise distance for a per-frame vector in a video. Smaller values indicate similar faces in consecutive frames of a generated video. However, the original ACD-C only signifies the face-identity-consistency between each pair of frames, lacking the information on general identity preservation. Therefore, we also use the **ACD-I** measure [53], the extension corresponding to the average of all $\mathcal{L}_2$ pairwise distances between each generated frame and the respective input frame.

4.3. Experimental Results

Transposed 3D *vs*. Transposed (1+2)D Convolution. Firstly, we compare video quality and training speed of our approach when using (i) transposed 3D convolutional filters only, and (ii) our transposed (1+2)D convolutional filters only, both having the same number of parameters for a fair comparison. The quantitative and qualitative results based on Weizmann dataset are presented in Table 2 and Figure 6.

Figure 6: **Sample generated frames** of ImaGINator with transposed 3D (top row) and transposed (1+2)D convolutions (bottom).

The results confirm that factorizing the transposed 3D convolutional filters into separate temporal and spatial components brings benefits: (i) an additional nonlinear rectification allows the model to represent more complex functions, (ii) optimization is facilitated, as transposed (1+2)D convolution blocks are easier to optimize than the full transposed 3D convolutional filters, and (iii) significant gains are yielded in both video quality and speed. Therefore, in the

	MUG		NATOPS		Weizmann		UvA-NEMO	
	SSIM/PSNR	FID	SSIM/PSNR	FID	SSIM/PSNR	FID	SSIM/PSNR	FID
VGAN [39]	0.28/14.54	74.72	0.72/20.09	167.71	0.29/15.78	127.31	0.21/13.43	30.01
MoCoGAN [37]	0.58/18.16	45.46	0.74/21.82	49.46	0.42/17.58	116.08	0.45/16.58	29.81
ImaGINator	**0.75/22.63**	**29.02**	**0.88/27.39**	**26.86**	**0.73/19.67**	**99.80**	**0.66/20.04**	**16.16**

Table 1: Evaluation of VGAN, MoCoGAN and proposed ImaGINator w.r.t. image quality (SSIM/PSNR) and video quality (FID).

Architecture	FID	Training time
Transposed 3D convolution	110.5	16.7s
Transposed (1+2)D convolution	**99.8**	**11.9s**

Table 2: **FID score and training time per epoch** of our approach with transposed 3D and transposed (1+2)D convolutions.

Methods	ACD-C	ACD-I
VGAN [39]	0.272	0.932
MoCoGAN [37]	0.158	0.904
ImaGINator	**0.141**	**0.431**
Reference	0.102	0.206

Table 3: **Evaluation of content consistency** of VGAN, MoCoGAN and proposed ImaGINator on the MUG dataset, represented by ACD-I and ACD-C scores.

Methods	Rater preference (%)
ImaGINator / MoCoGAN [37]	**83.32** / 16.68
ImaGINator / VGAN [39]	**85.43** / 14.57
MoCoGAN [37] / VGAN [39]	**70.85** / 29.15
ImaGINator / Real videos	20.82 / **79.18**

Table 4: **Subjective analysis.** Mean user preference of human raters comparing videos generated by the respective algorithms, as well as originated from all the datasets.

following evaluations we use our approach with the transposed (1+2)D convolution filters only.

ImaGINator. We proceed to compare our proposed ImaGINator to state-of-the-art video generation methods MoCoGAN and VGAN, both quantitatively and qualitatively. For the latter we report results pertained to a subjective analysis comparing the three methods. We then conduct an ablation study to prove the effectiveness of our proposed architecture, as well as to quantitatively evaluate the contribution of each part in our model.

Quantitative Analysis. For all methods, we sample 10 initial frames from each video sequence in each testing set. Both benchmark methods have been tuned with the best parameters on all training sets. All methods are trained to generate video sequences of 32 frames with an image size 64×64 pixels. Example generated frames of different methods are shown in Figure 5.

We firstly report reconstruction capabilities of our approach using SSIM and PSNR scores in Table 1. Our results show that the ImaGINator outperforms MoCoGAN and VGAN, w.r.t. SSIM and PSNR metrics, indicating that our proposed spatio-temporal fusion mechanism can well preserve the structure information of input image in the full generated video.

Then, we report FID scores for the three methods in Table 1. The ImaGINator achieves the lowest numbers on all four datasets, suggesting that videos generated by our method have the best temporal consistency and visual quality. This proves that modeling temporal consistency in higher spatial level can generate more realistic videos.

Then, we evaluate the content consistency for facial expression generation using ACD-C and ACD-I scores. Our results on the MUG dataset are presented in Table 3. The proposed ImaGINator outperforms both MoCoGAN and VGAN, on both ACD-C and ACD-I scores. The results confirm the ability of the proposed spatio-temporal fusion scheme to effectively preserve the appearance information in the generated videos.

Controllable Video Generation. We further conduct an experiment on the MUG and NATOPS datasets, where starting from the same image, we generate various videos associated to different labels (facial expressions / actions). Our results are presented in Figure 7. These results confirm the ability of our approach to generate new videos based on single images and category-labels.

Subjective Analysis. In addition, we conduct a subjective analysis, where we ask 30 human raters to pairwise compare videos generated by our approach with those generated by the state-of-the-art. We report the mean user preference in Table 4. We observe that human raters express a strong preference for the proposed framework over MoCoGAN (83.32% *vs*. 16.68%) and VGAN (85.43% *vs*. 14.57%), which is consistent with the above listed quantitative results. Further, we compare real videos from all the datasets with generated video sequences from our method. The human raters ranked 20.82% of videos from our ImaGINator as more realistic than real videos, which we find highly encouraging.

Ablation study. We here focus on showcasing the general effectiveness of our architecture, as well as the effectiveness related to each component of the proposed Generator.

Firstly, in the Generator G, we compare the performance of fully transposed 3D convolution with the proposed transposed (1+2)D convolution, and in the Discriminator D, we mainly focus on analyzing the usage of D_I. In addition, we compare each architecture with the model of the same architecture, but using an auxiliary classifier in D, similar

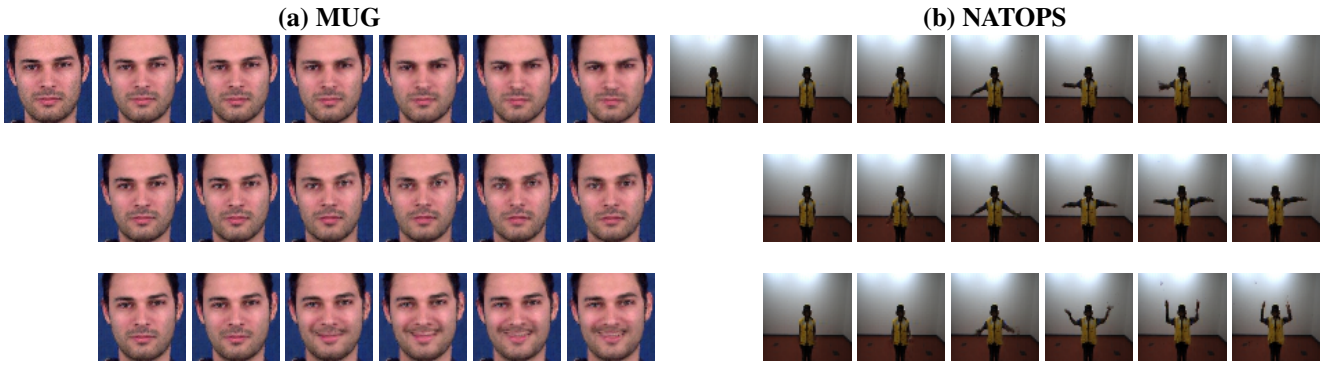


Figure 7: **Controllable video generation** in ImaGINator. Starting from the same image (top left for both datasets), we generate videos associated to different labels (remaining frames). In (a) MUG, from top to bottom the labels are set as *"fear"*, *"anger"* and *"happiness"*. In (b) NATOPS, from top to bottom the labels are set as *"all clear"*, *"fold winds"* and *"brakes on"*.

to ACGAN loss [29], which we refer as $D_V(ac)$. Our results are presented in Table 5. Our results show that given

Generator	Discriminator	MUG	NATOPS
3D	$D_V(ac)$	37.71	65.28
(1+2)D	$D_V(ac)$	32.57	52.43
3D	$D_V(ac), D_I$	33.08	57.65
(1+2)D	$D_V(ac), D_I$	29.91	48.41
3D	D_V	36.93	50.08
(1+2)D	D_V	29.80	40.57
3D	D_V, D_I	27.94	42.10
(1+2)D	D_V, D_I	**24.36**	**26.86**

Table 5: **Effectiveness of the proposed architecture**. We compare different architectures in both G and D to showcase the effectiveness of the proposed ImaGINator.

the same Discriminator, models using transposed (1+2)D convolution provide consistently lower video FID scores than models using transposed 3D convolution. The results confirm that our proposed transposed (1+2)D layer systematically improves video quality. Moreover, we show that adding D_I is beneficial, as well as that concatenating label vectors directly into spatio-temporal feature maps exceeds the performances of using auxiliary classifier in conditional video generation, see Table 5. This is especially true if the number of categories is large. A similar observation has been reported by Miyato and Koyama [28] in the context of conditional image generation.

Furthermore, we showcase that the spatio-temporal fusion contributes predominantly to video quality, see Table 6, and hence re-injecting spatial features and modeling temporal consistency in higher spatial level is an effective way to generate realistic videos. Finally, our results confirm that adding noise in the latent space is beneficial, as depicted in Table 6.

Further details w.r.t. our approach and experiments are presented in the supplementary material.

Architecture	MUG	NATOPS
ImaGINator, w/o ST fusion	46.02	62.89
ImaGINator, w/o (1+2)D	27.94	42.10
ImaGINator, w/o noise	32.38	32.05
ImaGINator	**24.36**	**26.86**

Table 6: **Contribution of main components in** G. We evaluate the ablation of spatio-temporal fusion, transposed (1+2)D convolution, as well as noise vector.

5. Conclusions

We have presented a novel conditional spatio-temporal GAN, namely ImaGINator, endowed with the ability to effectively generate videos based on a single image, a condition (label of a facial expression or action) and noise. Specifically, we focus on the settings, where we generate videos representing facial expressions and human actions, in which the human appearance is determined by a single input image, and the facial expression or human action is determined by a category-label, *e.g.* 'smile'. Our ImaGINator incorporates (a) a novel spatio-temporal fusion scheme, which generates dynamic motion, while retaining appearance throughout the full video sequence, and (b) a novel transposed (1+2)D convolution, factorizing the transposed 3D convolutional filters into separate transposed temporal and spatial components, which yields significant gains in video quality and speed. We have performed an extensive evaluation of our approach on 4 datasets, outperforming quantitatively and qualitatively the state-of-the-art video prediction methods. Our results have shown the efficiency of the ImaGINator in conditional image-to-video generation. Visualizations of the learned representation show that similar generation might be instrumental as augmented data, *e.g.* expression recognition in elderly subjects. We believe that video generation has the potential to affect many applications including simulations, forecasting, and representation learning.

References

[1] N. Aifanti, C. Papachristou, and A. Delopoulos. The mug facial expression database. In *WIAMIS*, 2010.

[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU School of Computer Science, 2016.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[5] E. Denton and R. Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.

[6] E. L. Denton and v. Birodkar. Unsupervised Learning of Disentangled Representations from Video. In *NIPS*, 2017.

[7] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *ECCV*, 2012.

[8] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 2007.

[11] K. Hara, H. Kataoka, and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *CVPR*, 2018.

[12] M. Hoai and F. De la Torre. Max-margin early event detectors. *IJCV*, 2014.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 2017.

[14] Y. Jang, G. Kim, and Y. Song. Video Prediction with Appearance and Motion Conditions. In *ICML*, 2018.

[15] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, 2017.

[16] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[19] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014.

[20] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017.

[21] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, 2018.

[22] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, 2017.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[24] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. *ICCV*, 2017.

[25] M. Mathieu, C. Couprie, and Y. LeCun. Deep Multi-Scale Video Prediction Beyond Mean Square Error. In *ICLR*, 2016.

[26] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[28] T. Miyato and M. Koyama. cGANs with projection discriminator. In *ICLR*, 2018.

[29] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis With Auxiliary Classifier GANs. In *ICML*, 2017.

[30] S. L. Pintea, J. C. van Gemert, and A. W. Smeulders. Déja vu. In *ECCV*, 2014.

[31] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[32] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, and B. Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *ECCV*, 2018.

[33] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011.

[34] Y. Song, D. Demirdjian, and R. Davis. Tracking Body and Hands For Gesture Recognition: NATOPS Aircraft Handling Signals Database. In *FG*, 2011.

[35] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[37] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018.

[38] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.

[39] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, 2016.

[40] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, 2016.

[41] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014.

[42] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[44] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva. G3AN: This video does not exist. Disentangling motion and appearance for video generation. *arXiv preprint arXiv:1912.05523*, 2019.

[45] Y. Wang, A. Dantcheva, and F. Bremond. From attribute-labels to faces: face generation using a conditional generative adversarial network. In *ECCVW*, 2018.

[46] Y. Wang, A. Dantcheva, and F. Bremond. From attributes to faces: a conditional generative adversarial network for face generation. In *BIOSIG*, volume 17, 2018.

[47] N. Wichers, R. Villegas, D. Erhan, and H. Lee. Hierarchical long-term video prediction without supervision. In *ICML*, 2018.

[48] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016.

[49] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin. Pose guided human video generation. In *ECCV*, 2018.

[50] S. Yu, H. Han, S. Shan, A. Dantcheva, and X. Chen. Improving face sketch recognition via adversarial sketch-photo transformation. In *FG*, 2019.

[51] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *ECCV*, 2010.

[52] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*, 2018.

[53] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*, 2018.

[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, 2017.