Dynamics Transfer GAN: Generating Video by Transferring Arbitrary Temporal Dynamics from a Source Video to a Single Target Image

Wissam J. Baddar, Geonmo Gu, Sangmin Lee and Yong Man Ro
Image and Video Systems Lab., School of Electrical Engineering, KAIST, South Korea
{wisam.baddar,geonm,sangmin.lee,ymro}@kaist.ac.kr

Abstract

In this paper, we propose Dynamics Transfer GAN; a new method for generating video sequences based on generative adversarial learning. The spatial constructs of a generated video sequence are acquired from the target image. The dynamics of the generated video sequence are imported from a source video sequence, with arbitrary motion, and imposed onto the target image. To preserve the spatial construct of the target image, the appearance of the source video sequence is suppressed and only the dynamics are obtained before being imposed onto the target image. That is achieved using the proposed appearance suppressed dynamics feature. Moreover, the spatial and temporal consistencies of the generated video sequence are verified via two discriminator networks. One discriminator validates the fidelity of the generated frames appearance, while the other validates the dynamic consistency of the generated video sequence. Experiments have been conducted to verify the quality of the video sequences generated by the proposed method. The results verified that Dynamics Transfer GAN successfully transferred arbitrary dynamics of the source video sequence onto a target image when generating the output video sequence. The experimental results also showed that Dynamics Transfer GAN maintained the spatial constructs (appearance) of the target image while generating spatially and temporally consistent video sequences.

1. Introduction

The recent advances in generative models have influenced researches to investigate image synthesis. Generative models, particularly generative adversarial networks (GANs), have been utilized to generate images from random distributions [4, 6], or synthesize images by nonlinearly transforming a priming image to the synthesized image [8, 7, 25], or even synthesizing images from a source image domain to a different domain [30, 10, 33].

The progress towards better image generation has been witnessing an interesting surge [10, 33, 5, 9, 17, 27]. Extending the capacities of generative models to generate video sequences is the natural and inevitable progression. However, extending generative models to generate meaningful video sequences is a challenging task. Generating meaningful video sequences requires the generative model to understand the spatial constructs of the scene as well as the temporal dynamics that drive the scene motions. In addition, the generative model should be able to reconstruct temporal variations with variable sequence lengths. In many cases, the dynamic motion could be non-rigid or cause shape transformation of the underlying spatial construct. As such, many aforementioned aspects could hamper the effectiveness of generative models to generate videos.

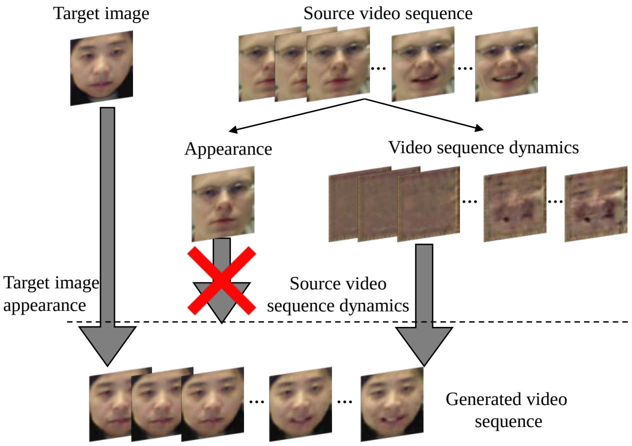


Figure 1: Proposed Dynamics Transfer GAN. Given a single target image and a video sequence with arbitrary temporal dynamics, the proposed Dynamics Transfer GAN generates a synthesized video sequence with the dynamics of the source video onto the appearance of the target image.

Due to the challenges mentioned above, some research efforts have tried to simplify the problem by limiting the generation to predict a few future frames of a given video sequence [3, 20, 15, 16, 28, 14, 13]. In these works, combinations of 3D convolutions and recurrent neural networks (RNNs) and convolutional long-short-term-memory (LSTM) were investigated to predict future frames. While many of these methods have shown impressive and promising results, predicting a few future frames is considered as a conditional image generation problem, which is different from the video generation [26].

The authors in [29] proposed an extension to GANs that generates videos with scene dynamics. The generator was composed of two streams to model the scene as a combination of a foreground and a background. 3D convolutions were utilized to perform spatio-temporal discriminators that criticize the generated sequences. A similar two-stream generator one spatio-temporal discriminator approach was proposed in [22]. Both [29, 22] could not model variable length video sequences, and could not generate long sequences. The authors of [26] separated the sampling procedure for input distribution into samples from a content subspace and a motion subspace to generate variable length sequences. The works in [26, 29, 22] showed that GANs could be extended to generate videos. However, the spatio-temporal discriminator was performed using 3D convolutions of fixed size, which meant that the spatio-temporal consistency of the generated videos could be limitedly verified at a fixed small sequence size. Moreover, the motion was coupled with the spatial construct in the spatio-temporal encoding process, which could limit the ability to generate dynamics at the desired spatial appearance.

In this paper, we propose a new video generation method named as Dynamics Transfer GAN. The proposed video generation is primed with a target image. The video sequence is generated by transferring the dynamics of arbitrary motion from a source video sequence onto the target image. The main contributions of the proposed method are summarized as follows:

1. We propose a new video sequence generation method from a single target image. The target image holds the spatial construct of the generated video sequence. The dynamics of the generated video are imported from an arbitrary motion of a source video sequence. The proposed method maintains the spatial appearance of the target image while importing the dynamics from a source video. To that end, we propose new appearance suppressed dynamics feature, which suppresses the spatial appearance of the source video while maintaining the temporal dynamics of source sequence. The proposed dynamics feature is devised so that the effect of spatial appearance is suppressed in the spatio-temporal encoding with a RNN. Thus, in video sequence generation, the source video dynamics are imported and imposed onto the target image.

2. The proposed Dynamics Transfer GAN is designed with the goal of generating variable length video sequences that extend in time (i.e., no limitation on the sequence length). To that end, we design a generator network with two discriminator networks. One discriminator investigates the fidelity of the generated frames (spatial discriminator). The other discriminator investigates the integrity of the generated sequence as a whole sequence (dynamics discriminator). In longer sequences, it could be expected that the dynamics discriminator focuses on the tailing parts of a video. To continuously maintain the spatial and dynamic fidelity of the generated video, additional objective terms were added to the training of the generator network. As a result, the proposed method generates videos with realistic spatial structure of the target image and temporal dynamics that mimic the source video sequence.

3. We provide a visualization of the imported dynamics from the source video. The dynamics visualization helps in understanding how the generative parts of the network perceive the input video sequence dynamics. Moreover, the visualization demonstrates that the proposed dynamics feature suppresses the source video appearance and only encodes the dynamics of the source video.

2. Related Work

2.1. Generative Adversarial Networks

Generative adversarial networks (GANs) have been proposed in [4] as a 2-player zero-sum game problem consisting of two networks: a generator network and a discriminator network. The generator network $(G : \mathbb{R}^K \rightarrow \mathbb{R}^M)$ tries to generate a sample $(\widehat{\mathbf{x}} \in \mathbb{R}^M)$ which mimics a sample in a given dataset $(\mathbf{x} \in \mathbb{R}^M)$. As an input, the generator network receives a latent variable $(\mathbf{z} \in \mathbb{R}^K)$, which is randomly sampled from a given distribution $p_{\mathbf{z}}$. Different distributions have been proposed to model the distribution of the latent variable $p_{\mathbf{z}}$, such as a Gaussian model [4], a mixture of Gaussians model [6] or even in the form of dropout to an input image [10]. On the other hand, the goal of the discriminator $(D : \mathbb{R}^M \rightarrow [0, 1])$ is to investigate the fidelity of the sample, and try to distinguish whether the given sample is real (ground truth sample) or fake (generated sample).

Training GANs can be achieved by simultaneously training the generator (G) and discriminator (D) networks with a non-cooperative game; i.e., the discriminator wins when it correctly distinguishes fake samples from real samples, while the generator wins when it generates samples that can fool the discriminator. Explicitly, the training of G and D is achieved by solving the minimax game with the value function:

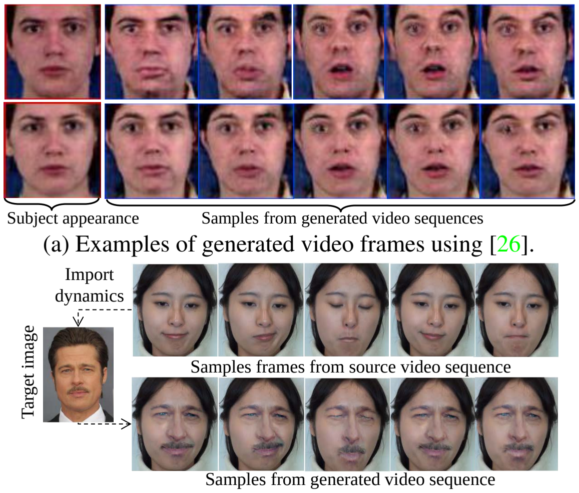
$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log{(1 - D(\mathbf{z}))}]. \tag{1}$$

2.2. Video Generation with GANs

Extending GAN to video generation is an instinctive progression from GAN for image generation. However, a few methods have tried generating complete video sequences [26, 29, 22, 18]. The authors of both [29, 22] have proposed a two-stream generator and one spatio-temporal discriminator approach. The authors in [29] assumed that a video sequence was constructed of a foreground and background and they separated generators accordingly. In [22], the video sequence was modeled by an image stream and a temporal stream. A limitation of these works is that the generator could limitedly generate fixed short-length video sequences.

The authors in [26] proposed MoCoGAN, a GAN for generating video sequences without a priming image. MocoGAN divided input distribution into content subspace and motion subspace. The sampling from the content subspace was accomplished by sampling from a Gaussian distribution. The sampling from the motion subspace was performed using an RNN. As such a content discriminator and motion discriminator were developed. MoCoGAN could generate sequences with variable lengths. However, the motion discriminator was limited to handle a fixed number of frames. This meant that the motion consistency of the generated videos was limitedly verified on a limited number of frames. Figure 2a shows examples of frames generated via MoCoGAN [26], in which the content of the generated video sequences was set to different subject appearances, while the motion was fixed to the same expression. As shown in the figure, the appearance of the generated video sequences is quite similar although the content (spatial construct) was different. In fact, in both cases, the subject identity of generated frames was fairly changed.

The authors in [18] proposed importing the dynamics from a source video sequence to a target image. However, this method resulted in severe disruption in the appearance of the target image. Figure 2b shows examples of frames taken from generated video sequences using [18]. It is clear from the figure that the method in [18] could capture the dynamics of the source video sequence. However, the spatial construct of the target image is severely damaged. The generated sequences follow the facial structure of the source sequence and append textural features of the target image onto it. In our proposed method, we intend to transfer the dynamics from the source video to the generated video while maintaining the appearance of the target image.


Subject appearance
Samples from generated video sequences
Import dynamics
Target image
Samples frames from source video sequence
Samples from generated video sequence

(a) Examples of generated video frames using [26].

(b) Examples of generated video frames using [18].

Figure 2: Examples of severe spatial-construct artifacts in previously proposed video generation methods.

3. The Proposed Dynamics Transfer GAN

Given an image $\mathbf{x}$ and a source video sequence $\mathbf{Y} = [\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_t, ..., \mathbf{y}_\mathrm{T}]$ with a frame $\mathbf{y}_i$ and sequence length T, the objective of the proposed method is to import the temporal dynamics from a source video and impose the dynamics on the input target image $\mathbf{x}$. As a result, the generator should generate a video sequence $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_0, \widehat{\mathbf{y}}_1, ...\widehat{\mathbf{y}}_t, ..., \widehat{\mathbf{y}}_\mathrm{T}]$ of length T and generated frame $\widehat{\mathbf{y}}_i$. The generated video sequence is supposed to possess the appearance of target image $\mathbf{x}$ and the dynamic of source sequence $\mathbf{Y}$. In the following subsections, we detail the proposed Dynamics Transfer GAN. First, we detail the proposed method for obtaining the input of Dynamics Transfer GAN. Then, the proposed GAN network structure and the training procedure with the proposed objective terms are explained.

3.1. Input of Dynamics Transfer GAN

The input of a video generative model could be represented as a vector of T samples in the latent space denoted as ($\mathbf{Z} = [\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_t, ..., \mathbf{z}_\mathrm{T}]$) [26]. Each sample $\mathbf{z}_t$ in $\mathbf{Z}$ represents a frame at time t. By traversing the samples of the vector $\mathbf{Z}$, we can explore the temporal path in which the video sequence traverses through. At any point in time t, $\mathbf{z}_t$ is decomposed into a spatial representation ($\mathbf{z}_t^{(s)}$) and a temporal dynamics representation ($\mathbf{z}_t^{(d)}$).

In this paper, we fix the spatial representation as the target image($\mathbf{z}_t^{(s)} = \mathbf{x}$), such that the spatial appearance follows the target image appearance. Note that when generating the spatial representation for a sample $\mathbf{z}_t^{(s)}$, instead of adding random noise to the target image $\mathbf{x}$, the noise is provided in the form of dropout applied on several layers of the generator as described in [10]. The dynamic representation, on the other hand, is attained from an appearance suppressed dynamics feature obtained using a pre-trained RNN [12]). The role of the RNN is to obtain a spatio-temporal representation of each frame in the source sequence. This allows us to represent dynamics at each frame as a sample point in the generator input space. Further, the RNN facilitates generating arbitrary length videos. In this paper,

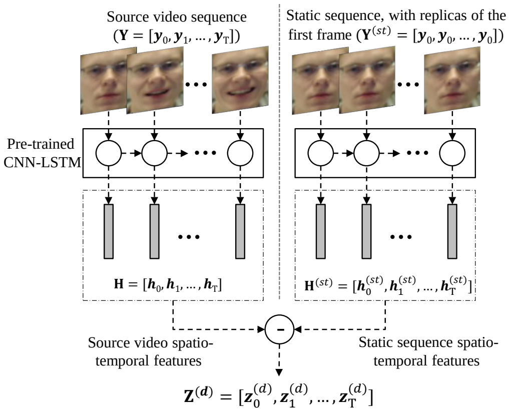

Source video sequence
$(\mathbf{Y} = [\boldsymbol{y}_0, \boldsymbol{y}_1, \dots, \boldsymbol{y}_T])$
Static sequence, with replicas of the first frame $(\mathbf{Y}^{(st)} = [\boldsymbol{y}_0, \boldsymbol{y}_0, \dots, \boldsymbol{y}_0])$
Pre-trained CNN-LSTM
$\mathbf{H} = [\boldsymbol{h}_0, \boldsymbol{h}_1, \dots, \boldsymbol{h}_T]$
$\mathbf{H}^{(st)} = [\boldsymbol{h}_0^{(st)}, \boldsymbol{h}_1^{(st)}, \dots, \boldsymbol{h}_T^{(st)}]$
-
Source video spatio-temporal features
Static sequence spatio-temporal features
$\mathbf{Z}^{(\boldsymbol{d})} = [\boldsymbol{z}_0^{(d)}, \boldsymbol{z}_1^{(d)}, \dots, \boldsymbol{z}_T^{(d)}]$

Figure 3: Appearance suppressed dynamics feature encoder for the input of Dynamics Transfer GAN.

when imposing the dynamics of the source video sequence to the target image, we are interested in imposing the dynamics while maintaining the target image appearance. We devise appearance suppressed dynamics feature to eliminate the effect of spatial encoding of the source video in the pre-trained RNN.

Figure 3 details the proposed method to obtain the appearance suppressed dynamics feature $\mathbf{Z}^{(d)}$. As shown in the figure, from the source video sequence $\mathbf{Y}$, a static sequence $\mathbf{Y}^{(st)} = [\mathbf{y}_0, \mathbf{y}_0, ..., \mathbf{y}_0]$ is generated by replicating the first frame of the source video sequence T times. Both the source video sequence $\mathbf{Y}$ and the static sequence $\mathbf{Y}^{(st)}$ are fed into the pre-trained RNN to generate the source video latent spatio-temporal features $\mathbf{H}$ and the static sequence latent spatio-temporal features $\mathbf{H}^{(st)}$, respectively. Since $\mathbf{Y}^{(st)}$ is T replicas of the same image, the RNN only encodes the spatial appearance in $\mathbf{H}^{(st)}$ rather than temporal features. Thus, by subtracting $\mathbf{H}^{(st)}$ from $\mathbf{H}$, the spatial appearance of the source video is suppressed and the dynamics of the source video are disentangled. Namely, the appearance suppressed dynamics feature can be obtained by $\mathbf{Z}^{(d)} = \mathbf{H} - \mathbf{H}^{(st)}$(see to the visualization of the appearance suppressed dynamics feature in Figure 7).

3.2. The Proposed Dynamics Transfer GAN

Figure 4 shows an overview of the proposed Dynamics Transfer GAN. The proposed Dynamics Transfer GAN is constructed of five main components: appearance suppressed dynamics feature encoder (A), dynamics channel embedder (F), generator network (G), spatial discriminator network (D_s) and a dynamics discriminator network (D_d).

The appearance suppressed dynamics feature encoder (mentioned in section 3.1) encodes the temporal dynamics feature of a source video. In this paper, to encode the dynamics effectively, the RNN model parameters employed from [12] are frozen during the training stage of the proposed Dynamics Transfer GAN.

The proposed Dynamics Transfer GAN generates a video sequence by synthesizing a sequence of frames. Hence both the spatial representation $(\mathbf{z}_t^{(s)})$ and the dynamic representation $(\mathbf{z}_t^{(d)})$ described in section 3.1 are fed into the generator at each frame. To combine the spatial representation $(\mathbf{z}_t^{(s)})$ with the dynamic representation $(\mathbf{z}_t^{(d)})$ at a time t, we embed the dynamics feature representation $(\mathbf{z}_t^{(d)})$ into a feature channel by using the dynamics feature embedder network (F). The embedded feature channel is concatenated with the target image $(\mathbf{z}_t^{(s)} = \mathbf{x})$ and fed to the generator. Note that we also use dropout on the dynamics feature embedder network to generate noise to the dynamics representation.

To maintain the target image appearance in the synthesized video, the generator network structure should be able to preserve the target image appearance. Many previous works in constructing a fine detailed image have utilized some form of encoder-decoder networks [11, 19, 31]. In this paper, we employ a U-net network structure for the generator network which could preserve details in image generator network [10, 21]. The generator input is the concatenation of the spatial representation $(\mathbf{Z}^{(s)})$ and the embedded dynamics feature representation channel $F(\mathbf{Z}^{(d)})$. The generator allows for a variable length video sequences $(\mathbf{Y})$ to be fed into it, so that it could generate a variable length sequence $(\widehat{\mathbf{Y}})$.

To criticize and discriminate the generated images, two discriminator networks are devised: spatial discriminator network (D_s) and dynamics discriminator network (D_d). The goal of the spatial discriminator is to check the fidelity of each generated frame, and try to distinguish real frames from generated frames. The spatial discriminator network structure is fairly straightforward. It is constructed of a stack of convolutional networks and an output layer for discriminating whether each frame is a real frame or a generated (fake) frame.

The purpose behind the dynamics discriminator is to distinguish if the dynamics of the generated sequence represent realistic dynamics or fake dynamics. The appearance suppressed dynamic feature for the generated sequence $(\widehat{\mathbf{Y}})$ is obtained by the appearance suppressed dynamics feature encoder (A). Similar to details described in Figure 3, to suppress the effect of appearance in the generated sequence, the RNN is utilized to obtain the generated static-latent features $(\widehat{\mathbf{H}}^{(st)})$ from a static sequence $\widehat{\mathbf{Y}}^{(st)} = [\widehat{\mathbf{y}}_0, \widehat{\mathbf{y}}_0, ..., \widehat{\mathbf{y}}_0]$. Accordingly, the generated appearance suppressed dynamics feature can be obtained by $(\widehat{\mathbf{Z}}^{(d)} = \widehat{\mathbf{H}} - \widehat{\mathbf{H}}^{(st)})$. At the dynamics discriminator, the realistic dynamics of a sequence is donated by $(\mathbf{Z}^{(d)})$ and fake (generated) sequence dynamics is denoted by $(\widehat{\mathbf{Z}}^{(d)})$.

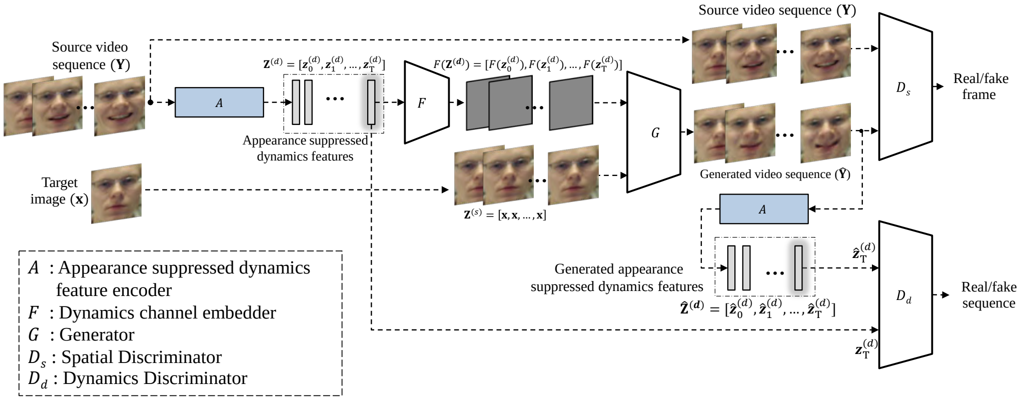

Source video sequence ($\mathbf{Y}$)
Source video sequence ($\mathbf{Y}$)
A
$\mathbf{Z}^{(d)} = [\mathbf{z}_0^{(d)}, \mathbf{z}_1^{(d)}, \dots, \mathbf{z}_{\mathrm{T}}^{(d)}]$
Appearance suppressed dynamics features
F
$F(\mathbf{Z}^{(d)}) = [F(\mathbf{z}_0^{(d)}), F(\mathbf{z}_1^{(d)}), \dots, F(\mathbf{z}_{\mathrm{T}}^{(d)})]$
G
D_s
Real/fake frame
Target image ($\mathbf{x}$)
$\mathbf{Z}^{(s)} = [\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}]$
Generated video sequence ($\hat{\mathbf{Y}}$)
A
Generated appearance suppressed dynamics features
$\hat{\mathbf{Z}}^{(d)} = [\hat{\mathbf{z}}_0^{(d)}, \hat{\mathbf{z}}_1^{(d)}, \dots, \hat{\mathbf{z}}_{\mathrm{T}}^{(d)}]$
$\hat{\mathbf{z}}_{\mathrm{T}}^{(d)}$
$\mathbf{z}_{\mathrm{T}}^{(d)}$
D_d
Real/fake sequence
A : Appearance suppressed dynamics feature encoder
F : Dynamics channel embedder
G : Generator
D_s : Spatial Discriminator
D_d : Dynamics Discriminator

Figure 4: Overview of the proposed Dynamics Transfer GAN

It should be noted that the dynamics discriminator and the spatial discriminator deal with the generated video differently. The spatial discriminator views generated video as a sequence of frames so that each frame represents a sample in the input space of spatial discriminator. The dynamics discriminator views the generated video sequence as a sample in the input space of dynamics discriminator. Note that a video sequence could have a variable length. To allow the dynamics discriminator to deal with the whole sequence as a sample point regardless of the sequence length, the input size of the dynamics discriminator should not be affected by the length of the sequence. Therefore, we only utilize the appearance suppressed feature ($\widehat{\mathbf{z}}_{\mathrm{T}}^{(d)}$) at time T. Due to the RNN properties, ($\widehat{\mathbf{z}}_{\mathrm{T}}^{(d)}$) represents the dynamics of full sequence, i.e., from the beginning until time T.

3.3. Training the Proposed Dynamics Transfer GAN

The training of the Dynamics Transfer GAN is a 2-player zero-sum game problem. Specifically, in this paper, the generative part of the network is a group of networks, i.e., the dynamics channel embedder (F) and the generator network (G). The discriminative part of Dynamics Transfer GAN is both the spatial discriminator (D_s) and the dynamics discriminator (D_d) networks. Note that the RNN parameters in the appearance suppressed dynamics feature encoder (A) are frozen. Thus, they are not included in the gradient update process. Explicitly, the training of F, G, D_s, and D_d is achieved by solving the minimax problem with the value function:

$$\begin{aligned}
&\min_{F,G}\max_{D_s,D_d} \mathcal{L}(F,G,D_s,D_d) = \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}[\log(D_s(\mathbf{Y})] \\
&+ \mathbb{E}_{\mathbf{z}^{(d)}\sim p_{\mathbf{z}}^{(d)}}[\log(D_d(\mathbf{z}_{\mathrm{T}}^{(d)}))] \\
&+ \mathbb{E}_{\mathbf{Z}^{(d)}\sim p_{\mathbf{Z}}^{(d)},\mathbf{Z}^{(s)}\sim p_{\mathbf{Z}}^{(s)}}[\log(1 - D_s(G(\mathbf{Z}^{(s)}, F(\mathbf{Z}^{(d)}))))] \\
&+ \mathbb{E}_{\widehat{\mathbf{z}}^{(d)}\sim p_{\widehat{\mathbf{z}}}^{(d)}}[\log(1 - D_d(\widehat{\mathbf{z}}_{\mathrm{T}}^{(d)}))]. \qquad (2)
\end{aligned}$$

Note that the back propagation can be performed independently on the discriminator networks (D_d, D_s). The generative networks (F, G) are updated after both discriminators (D_d, D_s). Therefore, in practice, the training of the discriminators (D_d, D_s) and that of the generative networks (F, G) are performed alternatively. In the first step, discriminators (D_d, D_s) are trained by maximizing the loss terms:

$$\begin{aligned}
&\mathcal{L}_{D_s}(F,G,D_s) = \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}[\log(D_s(\mathbf{Y})] \\
&+ \mathbb{E}_{\mathbf{Z}^{(d)}\sim p_{\mathbf{Z}^{(d)}},\mathbf{Z}^{(s)}\sim p_{\mathbf{Z}^{(s)}}}[\log(1 - D_s(G(\mathbf{Z}^{(s)}, F(\mathbf{Z}^{(d)}))))], \\
&\mathcal{L}_{D_d}(F,G,D_d) = \mathbb{E}_{\mathbf{z}^{(d)}\sim p_{\mathbf{z}^{(d)}}}[\log(D_d(\mathbf{z}_{\mathrm{T}}^{(d)}))] \\
&+ \mathbb{E}_{\widehat{\mathbf{z}}^{(d)}\sim p_{\widehat{\mathbf{z}}^{(d)}}}[\log(1 - D_d(\widehat{\mathbf{z}}_{\mathrm{T}}^{(d)}))]. \qquad (3)
\end{aligned}$$

In the second step, the generative parts of the network are trained by minimizing the adversarial loss:

$$\begin{aligned}
&\mathcal{L}_{G^{(A)}}(F,G,D_s,D_d) = \mathbb{E}_{\widehat{\mathbf{z}}^{(d)}\sim p_{\widehat{\mathbf{z}}^{(d)}}}[-\log(D_d(\widehat{\mathbf{z}}^{(d)}))] \\
&+ \mathbb{E}_{\mathbf{Z}^{(d)}\sim p_{\mathbf{Z}^{(d)}},\mathbf{Z}^{(s)}\sim p_{\mathbf{Z}^{(s)}}}[-\log(D_s(G(\mathbf{Z}^{(s)}, F(\mathbf{Z}^{(d)}))))]. \\
&\qquad (4)
\end{aligned}$$

Adding a reconstruction term to the generative networks could improve the quality of the generated images [10], We employ L1 frame reconstruction loss to improve the spatial reconstruction at each frame as follows:

$$\mathcal{L}_{G^{(s)}}(F,G) = \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}\left[\left\|\mathbf{Y} - G(\mathbf{Z}^{(s)}, F(\mathbf{Z}^{(d)}))\right\|_1\right]. \quad (5)$$

Finally, to maintain dynamic consistency even when the sequence is lengthy, we propose a reconstruction term on the appearance suppressed dynamics feature. Unlike the loss from the dynamics discriminator (which is calculated at the end of the sequence), this term makes sure that the generated sequence dynamics are maintained at each frame.

$$\mathcal{L}_{G^{(d)}}(F,G) = \mathbb{E}_{\mathbf{Z}^{(d)}\sim p_{\mathbf{Z}^{(d)}}}\left[\left\|\mathbf{Z}^{(d)} - \widehat{\mathbf{Z}}^{(d)}\right\|_1\right]. \quad (6)$$

By plugging in eqs. (5) and (6) in the second step of the training described in eq.(4), the generator training can be obtained by minimizing the loss:

$$\mathcal{L}_G(F, G, D_s, D_d) = \lambda_{G^{(A)}} \mathcal{L}_{G^{(A)}}(F, G, D_s, D_d) + \lambda_{G^{(s)}} \mathcal{L}_{G^{(s)}}(F, G) + \lambda_{G^{(d)}} \mathcal{L}_{G^{(d)}}(F, G), \quad (7)$$

where $\lambda_{G^{(A)}}$, $\lambda_{G^{(s)}}$, and $\lambda_{G^{(d)}}$ are hyper-parameters to control the generative loss term. $\lambda_{G^{(A)}}$ is the weight controlling the adversarial loss of the Dynamics Transfer GAN described in eq. (4), $\lambda_{G^{(s)}}$ is the weight controlling the frame reconstruction loss detailed in eq. (5), and $\lambda_{G^{(d)}}$ is the weight controlling the loss that maintains the dynamic consistency at each frame described in eq. (6).

4. Experiments

4.1. Experiment Setup

To verify the effectiveness of the proposed Dynamics Transfer GAN, experiments were conducted on the Oulu-CASIA dataset [32]. In the dataset, sequences of the six basic expressions (i.e., angry, disgust, fear, happy, sad, and surprise) were collected from 80 subjects under three illumination conditions. For the experiments, video sequences collected with a visible light camera under normal illumination conditions were used. A total of 480 video sequences were collected (6 video sequences per subject, 1 video sequence per expression). For each subject, the basic expression sequence was captured from a neutral face until the expression apex. In the experiments, the face region was detected and facial landmark detection was performed [2] on each frame. The face region was then automatically cropped and aligned based on the eye landmarks [24]. The training and implementation of the proposed Dynamics Transfer GAN have been conducted using TensorFlow [1].

4.2. Experiment 1: Visual Qualitative Results of Videos Generated with the Proposed Dynamics Transfer GAN

In experiment 1, we investigated the quality of the generated video sequences by using the proposed Dynamics Transfer GAN. Figure 5 shows examples of generated sequences to qualitatively assess the generated video sequence (more examples of generated video sequences can be found in the appendix, section 6.1). The figure shows the source video sequence, the target image and the corresponding generated video sequences. As seen in the figure, the generated videos are of arbitrary length, and the length of the sequence has no effect on the quality of the generated image. This is attributed to the dynamics discriminator, and the generator dynamics objective term. From the figure, it can also be seen that multiple videos could be generated from the same source video sequence. This was achieved by imposing the dynamics of the source video sequence onto different target images. The figure also presents an example of the same appearance with different dynamics (e.g., different expression sequence imposed on the same target image). This was accomplished by fixing the target image, while changing the source video sequence. These results show that the proposed method effectively transfers arbitrary dynamics of source video sequences to the generated video sequences. In addition, the appearance and the identity of the target image were preserved in the generated video sequences.

4.3. Experiment 2: Subjective Assessment of Videos Generated with the Proposed Dynamics Transfer GAN

It is known that quantitatively evaluating generative models, especially on generating visual outputs, is challenging [26]. It is also shown that all the popular metrics are subject to flaws [23]. Therefore, in experiment 2, we performed a subjective experiment in order to quantitatively evaluate the generated video sequences [26]. To that end, 10 participants took part of the experiment to evaluate the quality of the generated videos. The participants viewed a total of 240 generated video sequences. The video sequences were generated by transferring the dynamics from 24 source video sequences into 10 target images. The source sequences represented the 6 basic facial expressions (i.e., anger, disgust, fear, happiness, sadness and surprise) of 4 subjects. The target images were neutral expression images of 10 subjects. When evaluating the generated video sequences, the participants were guided to watch the generated video sequences along with the corresponding source video sequences and target images. After viewing each sequence, the subjects were asked to rate if the generated video is realistic or not. Table 1 shows the percentage of the video sequences that were rated as realistic. As can be seen from the results, more that 78.33% of the videos were rated as realistic. This percentage reflects that the proposed Dynamics Transfer GAN has generated video sequences with a reasonable quality.

In evaluating the quality of the generated videos, it is important to make sure that (1) the generated videos preserved the appearance of the target image, and that (2) the dynamic transitions in the video sequence should be smooth and consistent. To that end, each subject in the subjective assessment experiment was asked to rate the spatial consistency of the generated video sequence (i.e., the spatial appearance of the target image was preserved, and the spatial constructs in the generated frames was intact). Added to that, the subjects were asked to rate if the generated videos were temporally consistent (i.e., the dynamic transition between the frames was smooth). The results of the subjective evaluation are shown in Table 1. The results show that the proposed Dynamics Transfer GAN could generate temporally consistent sequences while preserving the appearance of the target image.

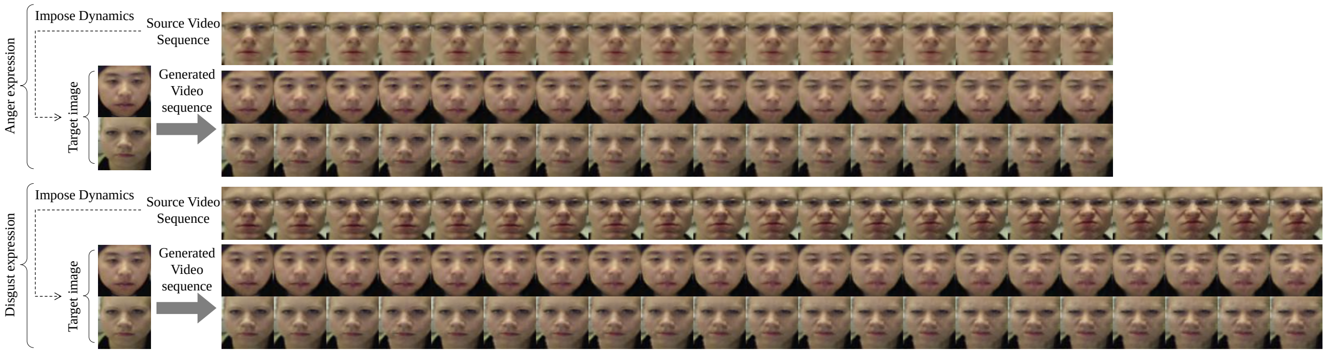
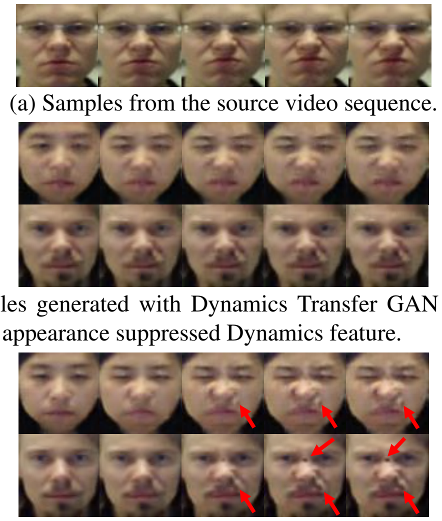


Figure 5: Example of sequences of arbitrary length generated by the proposed Dynamics Transfer GAN.

Table 1: Subjective quality assessment results for the generated video sequence

Question	Yes(%)	No(%)
Is this video realistic	78.33%	21.67%
Is the video temporally consistent	82.92%	17.08%
Is the video spatially consistent	79.17%	20.83%



(a) Samples from the source video sequence.

(b) Samples generated with Dynamics Transfer GAN using the proposed appearance suppressed Dynamics feature.

(c) Samples generated with Dynamics Transfer GAN using the CNN-LSTM features.

Figure 6: Comparison of video frames generated using Dynamic Transfer GAN with different source video sequence dynamic feature encodings.

4.4. Experiment 3: Comparison with Different Source Video Dynamic Feature Encodings

In this experiment, we investigated the quality of the generated video sequences with the Dynamics Transfer GAN by using different dynamic feature encoding. To that end, we generated video sequences using the proposed Dynamics Transfer GAN that utilizes the proposed appearance suppressed dynamics feature $(\mathbf{Z}^{(d)} = \mathbf{H} - \mathbf{H}^{(st)})$as detailed in Figure 3 . For comparison, different dynamic feature encoding was performed by replacing the appearance suppressed dynamics feature with the CNN-LSTM features [12] (the source video spatio-temporal feature in Figure 3 is used as input of the dynamic channel embedder (F) in Figure 4 , i.e., $\mathbf{Z}^{(d)} = \mathbf{H}$).

Figure 6 shows a number of example frames from the generated video sequences using the proposed Dynamics Transfer GAN. Figure 6a shows frames from the source video sequence. The images in Figure 6b were generated using the proposed Dynamic Transfer GAN with proposed appearance suppressed dynamics features $(\mathbf{Z}^{(d)} = \mathbf{H} - \mathbf{H}^{(st)})$. The images in Figure 6c were generated using the proposed Dynamic Transfer GAN with the CNN-LSTM features [12]. For more examples, please refer to the appendix section 6.2. By inspecting the location of the artifacts in the images generated using the CNN-LSTM features, we observed that artifacts could mainly occur at: (1) frames with larger dynamics (frames with intense expressions) and (2) locations where there was a deformation in the spatial construct of the source video frame (e.g., wrinkle locations). On the other hand, such artifacts were minimized when the Dynamics Transfer GAN utilized the appearance suppressed dynamics features. These results show that the proposed appearance suppressed dynamics feature could encode the dynamics of the source features more efficiently. As a result, the generated images have fewer artifacts compared to the model utilizing the CNN-LSTM features.

We further performed a subjective preference experiment, to quantitatively evaluate the effect of the source video sequence dynamic encoding method on the generated videos. To that end, 10 participants were summoned. This time, the subjects were presented with 240 pairs of video sequences. In each pair, one video was generated using the proposed appearance suppressed dynamics feature. The other video was generated using the CNN-LSTM features. The subjects were asked to state which video sequence they preferred, in terms of quality. Note that to avoid bias in the subject's decisions, the location of the presented pair of sequences was switched randomly. Table 2 shows that the videos generated using the appearance suppressed dy-

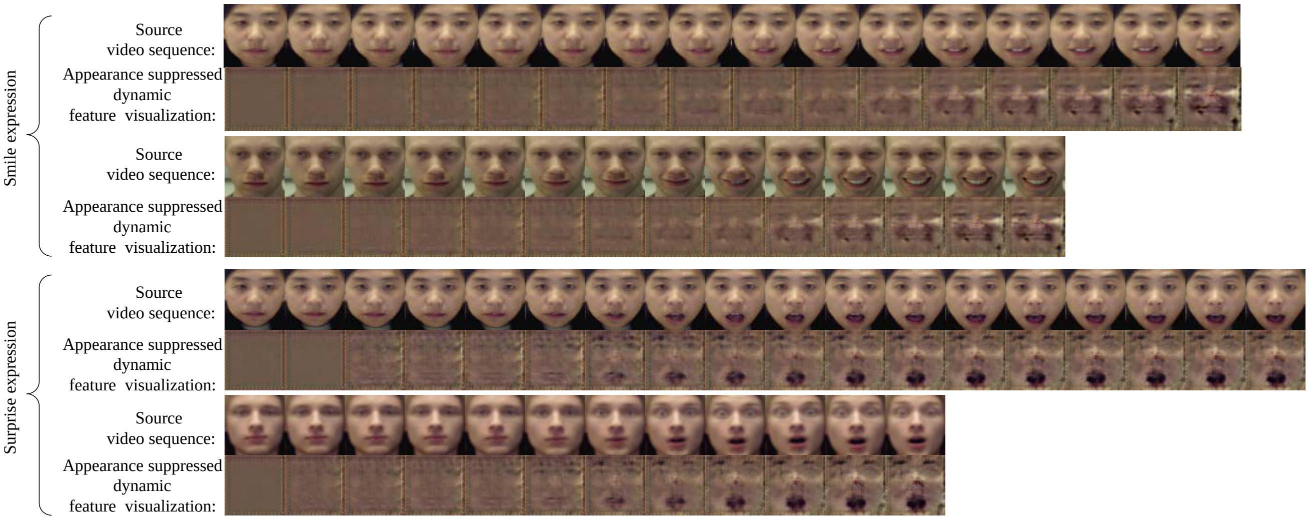

Smile expression
Source video sequence:
Appearance suppressed dynamic feature visualization:
Source video sequence:
Appearance suppressed dynamic feature visualization:
Surprise expression
Source video sequence:
Appearance suppressed dynamic feature visualization:
Source video sequence:
Appearance suppressed dynamic feature visualization:

Figure 7: Visualization of the appearance suppressed dynamic feature.

namics feature were overwhelmingly preferred by the test subjects. These results support the fact that the appearance suppressed dynamics feature is more efficient in encoding the dynamics of the source video sequence compared to the spatio-temporal features of the CNN-LSTM [12]. It also confirms the qualitative results that generated video sequence generated via the appearance suppressed dynamics feature are less prone to artifacts.

Table 2: Subject preference results for the generated videos (the proposed appearance suppressed dynamics features vs. CNN-LSTM features)

	CNN-LSTM features	**Appearance suppressed dynamics features**
Preference (%)	17.12%	82.88%

4.5. Experiment 4: Visualization of the Appearance Suppressed Dynamics Features

In experiment 4, we intend to visualize the appearance suppressed dynamics feature of different source video sequence. After the Dynamics Transfer GAN is trained, a video sequence is generated by importing the dynamics of a source video sequence, and imposing them on a target image. However, we wondered what would happen if the target image did not contain any spatial construct (i.e., an image with zero pixel values). This should provide a visualization for the appearance suppressed dynamics features of the source video. To test that hypothesis, video sequences were constructed with a target image of no spatial construct (image with zero pixel values). Examples of the resulting video sequences are shown in Figure 7. As shown in the figure, the generated video sequences share the same dynamics with the source video sequence. However, the appearance (identity of the subject) in the source sequence was removed. Another way to interpret these results is that the generator constructed a visualization of the appearance suppressed dynamics features (which was obtained from the source video sequence via the RNN). This visualization validates that the proposed appearance suppressed dynamics features suppress the spatial appearance of the source video sequence.

5. Conclusion

In this paper, we proposed a new video generation method based on GAN. The proposed method transfers arbitrary dynamics from a source video sequence to a target image. The spatial construct of a generated video sequence acquires the appearance of the target image. To achieve that, the appearance of the source video sequence was suppressed and only the dynamics of the source video were obtained before being imposed onto the target image. Therefore, an appearance suppressed dynamics feature was proposed to diminish the appearance of source video sequence while strictly encoding the dynamics. The appearance suppressed dynamics features utilized a pre-trained RNN network. Two discriminators were proposed: the spatial discriminator to validate the fidelity of the generated frames appearance and the dynamics discriminator to validate the continuity of the generated video dynamics.

Qualitative and subjective experiments have been conducted to verify the quality of the generated videos with the proposed Dynamics Transfer GAN. The results verified that the proposed method was able to impose the arbitrary dynamics of a source video sequence onto a target image, without distorting the spatial construct of that image. The experiments also verified that Dynamics Transfer GAN generates spatially and temporally consistent video sequences.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 6

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1859–1866. IEEE, 2014. 6

[3] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016. 1

[4] 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. 1, 2

[5] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of Machine Learning Research (PMLR)*, pages 1462–1471, 2015. 1

[6] S. Gurumurthy, R. K. Sarvadevabhatla, and V. B. Radhakrishnan. Deligan: Generative adversarial networks for diverse and limited data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 166–174, 2017. 1, 2

[7] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu. Real-time neural style transfer for videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 783–791, 2017. 1

[8] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 1

[9] D. J. Im, C. D. Kim, H. Jiang, and R. Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016. 1

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 1, 2, 3, 4, 5

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 4

[12] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 2017. 3, 4, 7, 8

[13] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual motion gan for future-flow embedded video prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1744–1752, 2017. 1

[14] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4463–4471, 2017. 1

[15] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[16] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[17] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug and play generative networks: Conditional iterative generation of images in latent space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4467–4477, 2017. 1

[18] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, and H. Li. Realistic dynamic facial textures from a single image using gans. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5429–5438, 2017. 3

[19] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 4

[20] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. In *International Conference on Learning Representations (ICLR): Workshop Track*, 2016. 1

[21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4

[22] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2830–2839, 2017. 2, 3

[23] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016. 6

[24] Y.-l. Tian. Evaluation of face resolution for expression analysis. In *Computer Vision and Pattern Recognition Workshop*, pages 82–82. IEEE, 2004. 6

[25] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4, pages 1415–1424, 2017. 1

[26] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*, 2017. 2, 3, 6

[27] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016. 1

[28] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[29] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 2, 3

[30] L. Wolf, Y. Taigman, and A. Polyak. Unsupervised creation of parameterized avatars. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1530–1538, 2017. 1

[31] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision (ECCV)*, pages 517–532. Springer, 2016. 4

[32] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 6

[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 1

6. Appendix

6.1. Additional Examples of Sequences Generated Using the Proposed Dynamics Transfer GAN

In the main paper (particularly section 3.2), we showed a few examples of generated video sequences using the proposed Dynamics Transfer GAN. In Figures 8,9,10,11,12 and 13, we show additional examples of different generated facial expression sequences. In each figure, two different source video sequences, three target images and the corresponding generated video sequences. Please refer to the video https://youtu.be/ppAUF1WVur8to watch playable examples of generated video sequences.

6.2. Additional Comparison Examples with Different Source Video Dynamic Feature Encodings

In the section 3.4 of the main paper, we showed some comparative examples of video frames generated using Dynamic Transfer GAN with different source video sequence dynamic feature encodings. In Figures 14, additional examples are shown with different subjects and facial expressions. Please refer to the video at https://youtu.be/ppAUF1WVur8 to watch playable examples of comparison example video sequences.

6.3. Examples of Generated Video Sequences Imported from Long Source Video Sequences

In the video https://youtu.be/ppAUF1WVur8, we provide examples of generated lengthy videos sequences (large number of frames). The dynamics of a generated video was imported from one long video sequence composed of the 6 basic expressions performed continuously. From the results shown in the video examples, the proposed Dynamics Transfer GAN generated video sequences with reasonable quality regardless of the lengthy sequences.

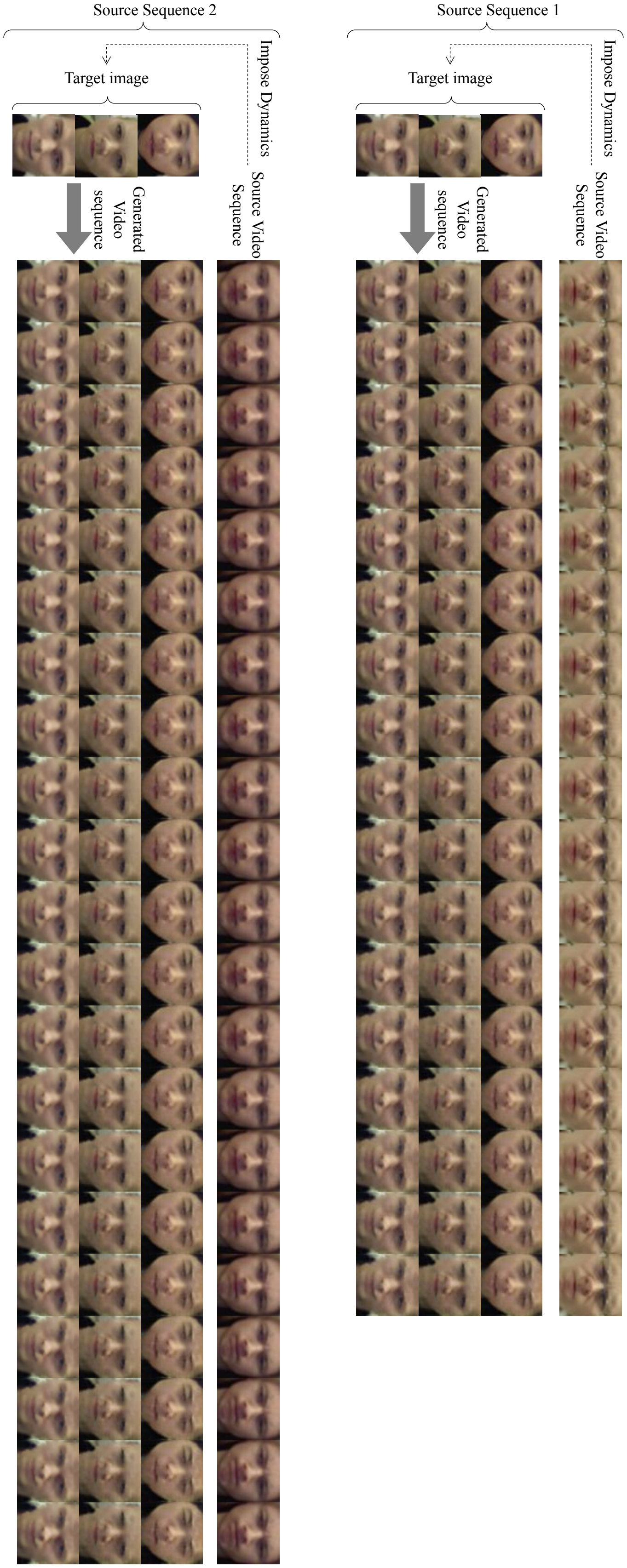


Figure 8: Examples of video sequences (anger expression) generated with the proposed Dynamics Transfer GAN.

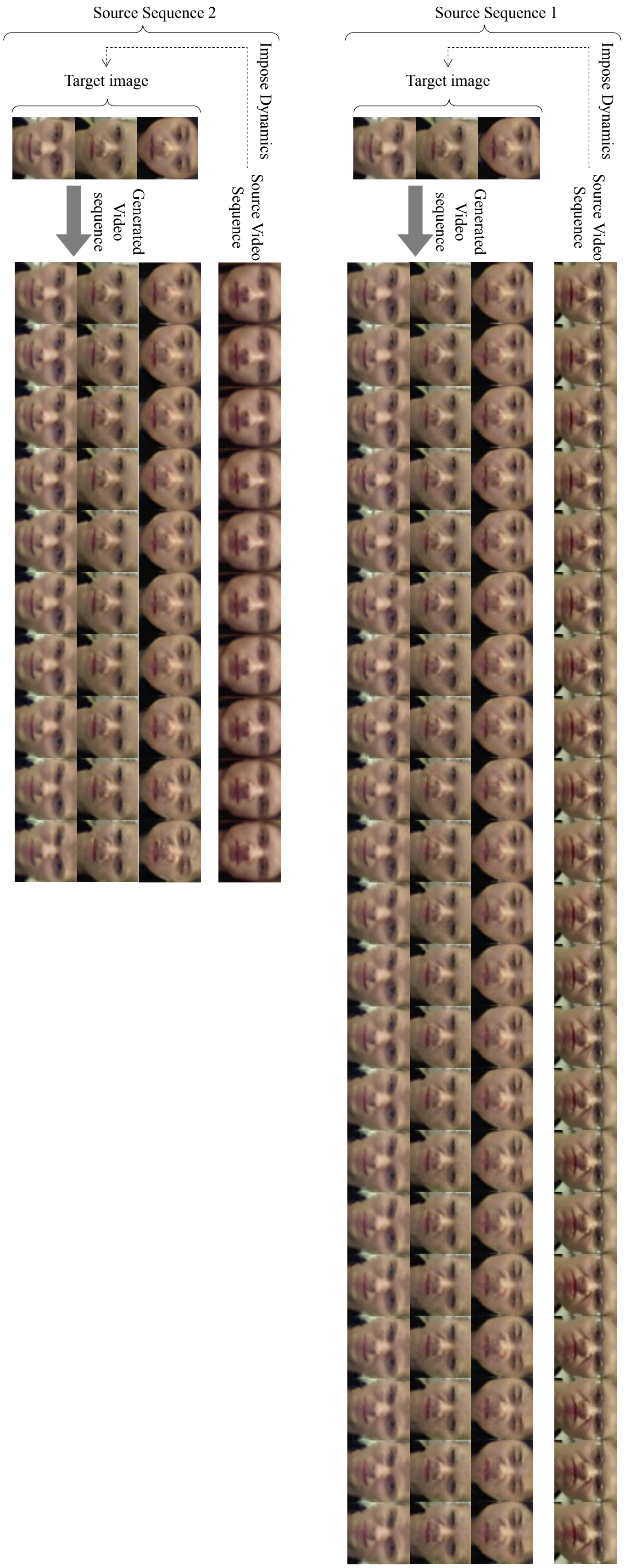


Figure 9: Examples of video sequences (disgust expression) generated with the proposed Dynamics Transfer GAN.

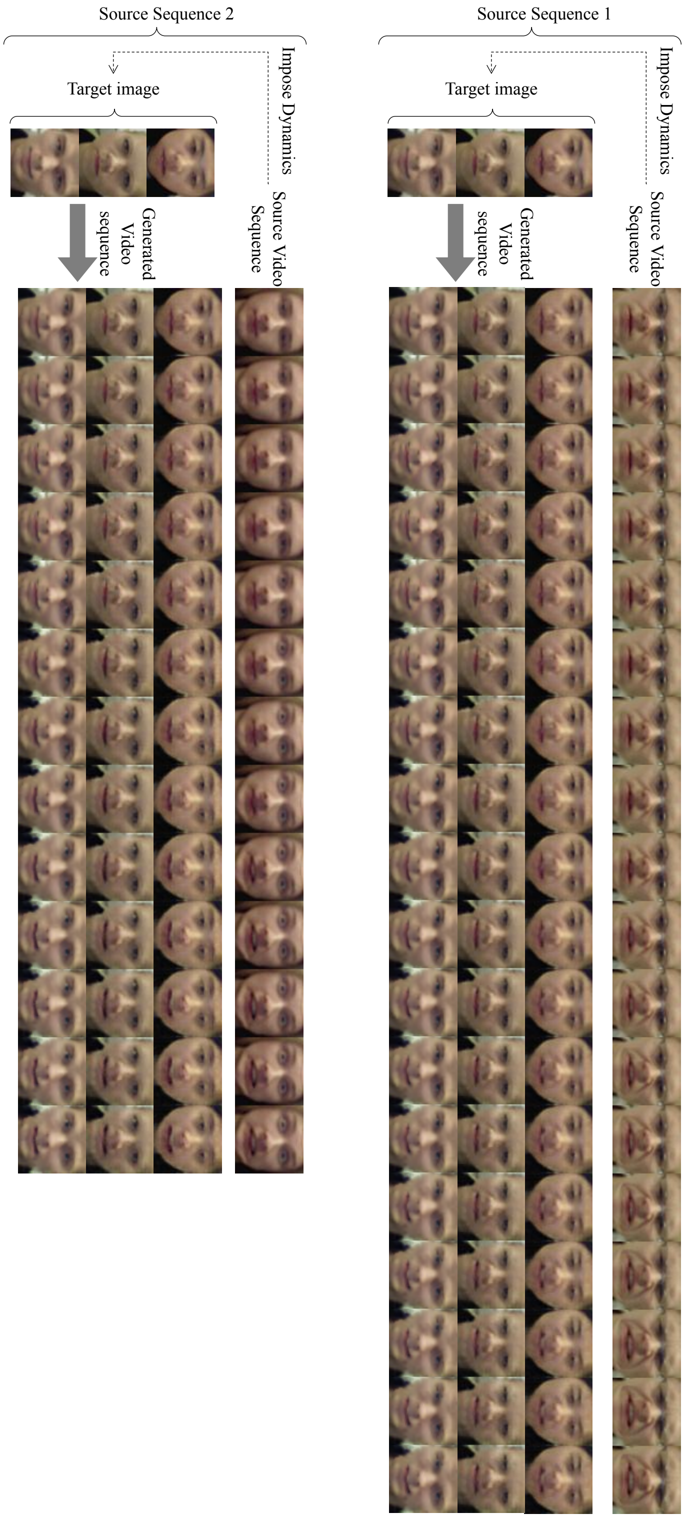


Figure 10: Examples of video sequences (fear expression) generated with the proposed Dynamics Transfer GAN.

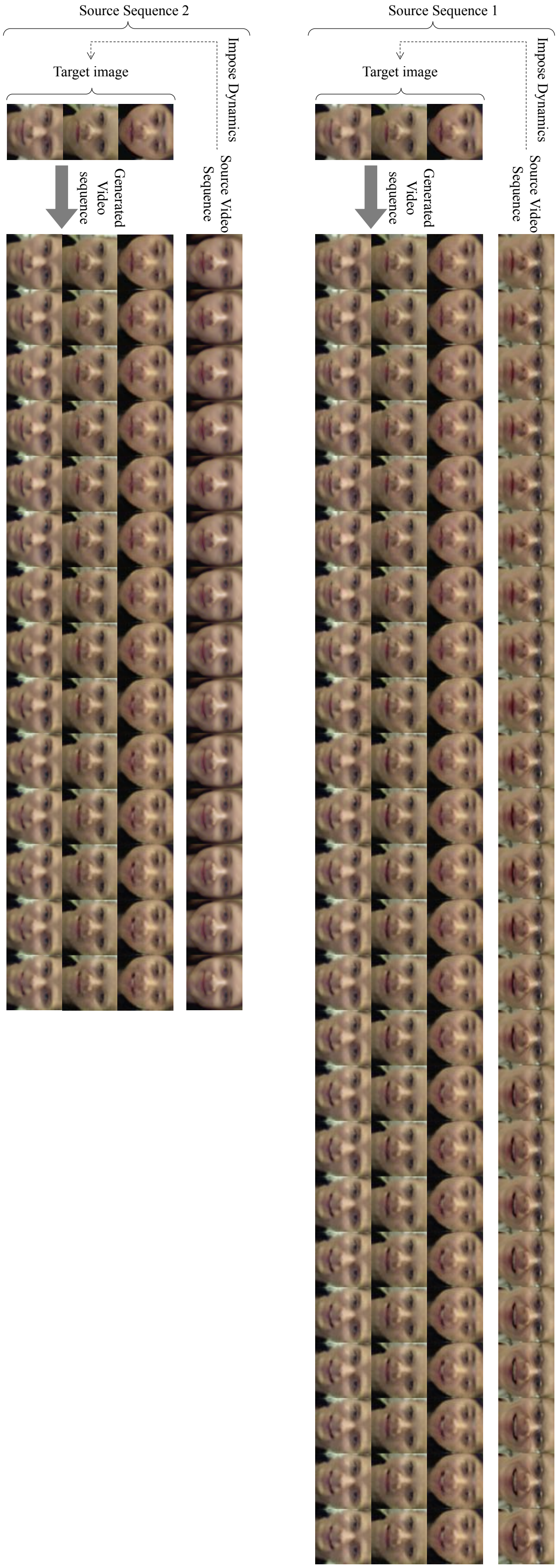


Figure 11: Examples of video sequences (happiness expression) generated with the proposed Dynamics Transfer GAN.

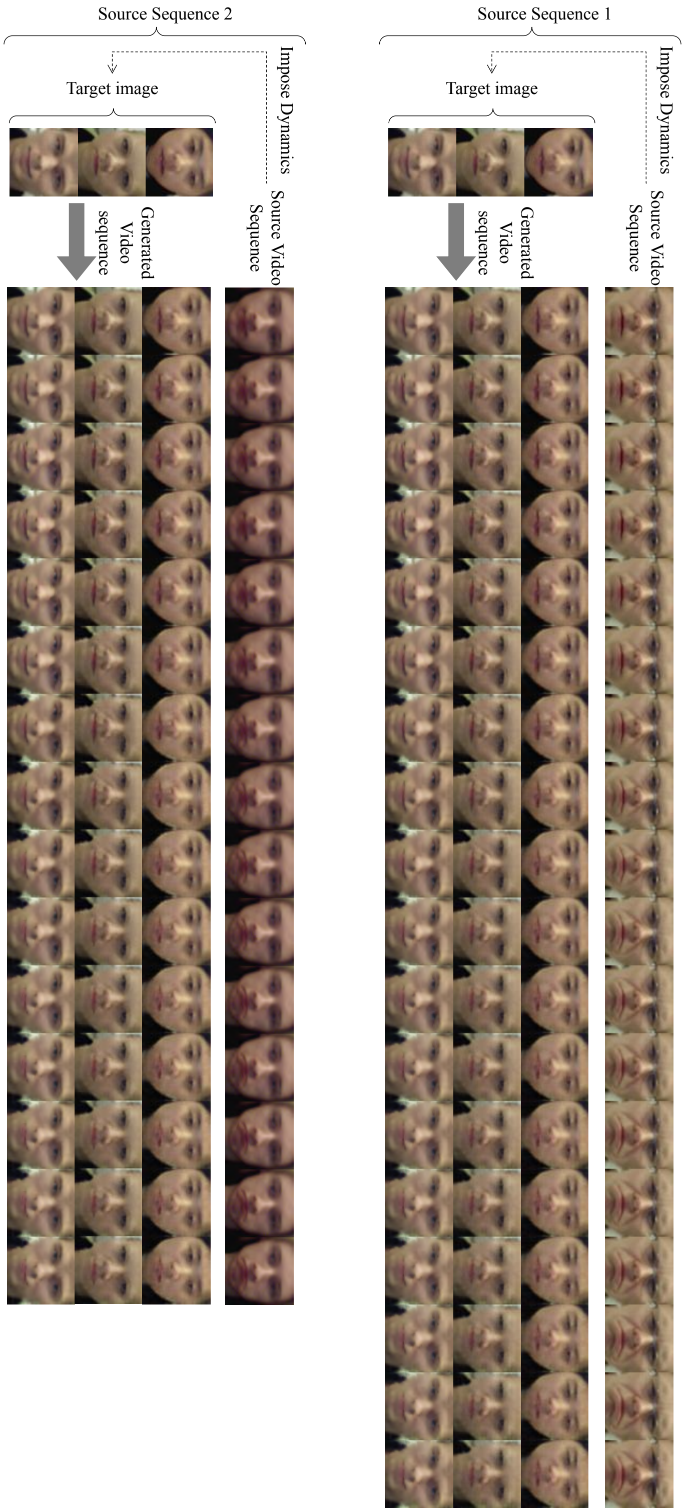


Figure 12: Examples of video sequences (sadness expression) generated with the proposed Dynamics Transfer GAN.

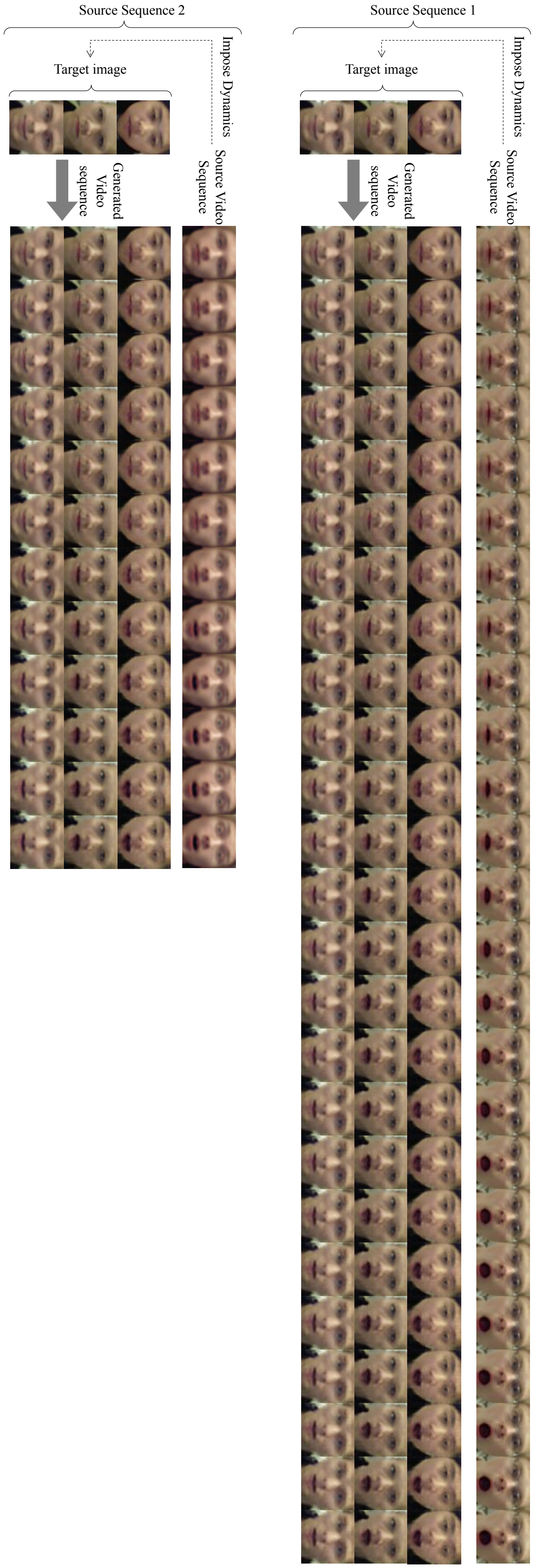
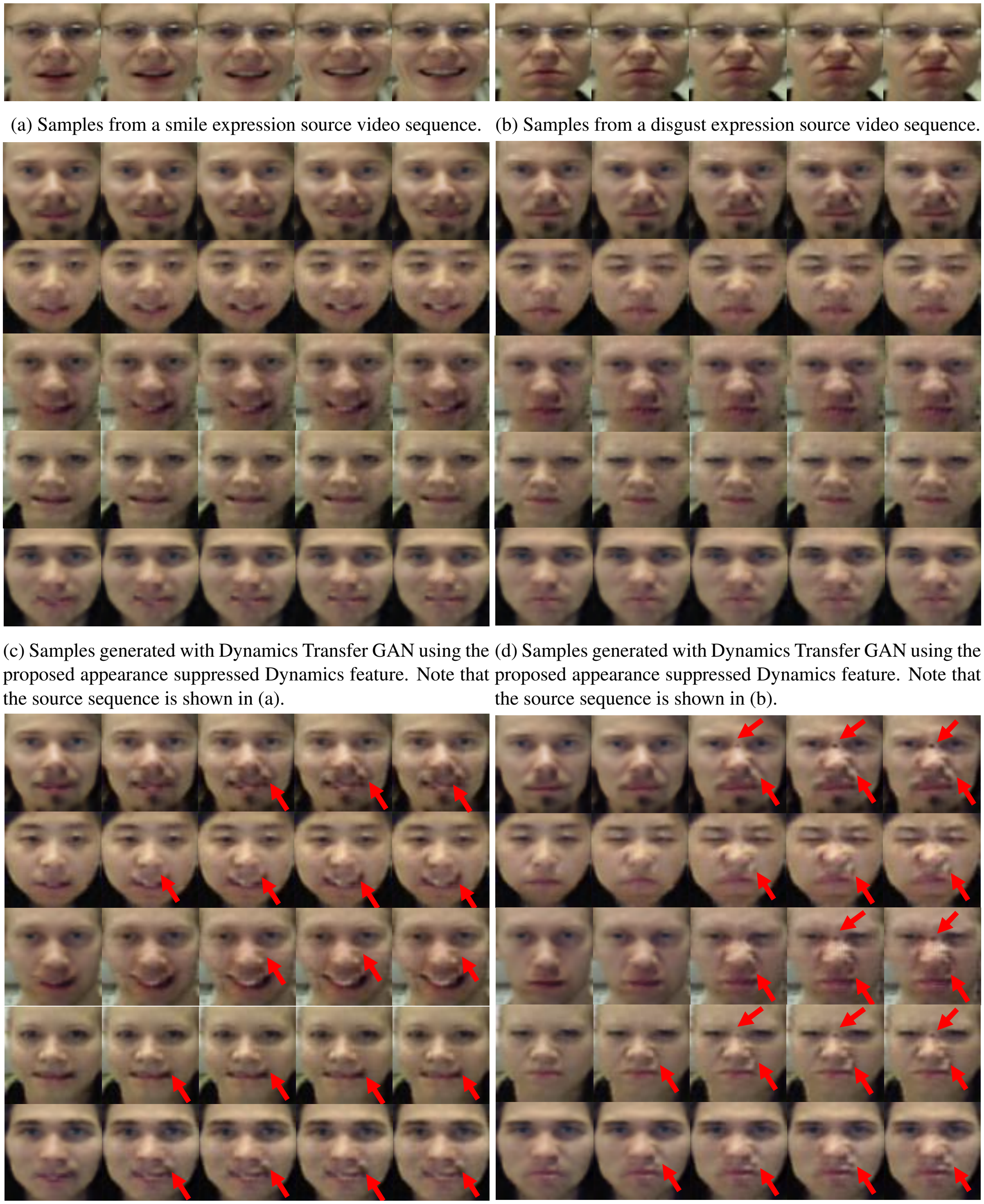


Figure 13: Examples of video sequences (surprise expression) generated with the proposed Dynamics Transfer GAN.



(a) Samples from a smile expression source video sequence.

(b) Samples from a disgust expression source video sequence.

(c) Samples generated with Dynamics Transfer GAN using the proposed appearance suppressed Dynamics feature. Note that the source sequence is shown in (a).

(d) Samples generated with Dynamics Transfer GAN using the proposed appearance suppressed Dynamics feature. Note that the source sequence is shown in (b).

(e) Samples generated using CNN-LSTM features. Note that the source sequence is shown in (a).

(f) Samples generated using CNN-LSTM features. Note that the source sequence is shown in (b).

Figure 14: Comparison of video frames generated using Dynamic Transfer GAN with different source video sequence dynamic feature encodings.