The Visual Computer
https://doi.org/10.1007/s00371-020-01913-6

ORIGINAL ARTICLE

Unsupervised video-to-video translation with preservation of frame modification tendency

Huajun Liu[1] · Chao Li[1] · Dian Lei[1] · Qing Zhu[2]

Abstract
Tremendous advances have been achieved in image translation with the employment of generative adversarial networks (GANs). With respect to video-to-video translation, similar idea has been leveraged by various researches, which may focus on the associations among relevant frames. However, the existing video-synthesis methods based on GANs do not make full exploitation of the spatial–temporal information in videos, especially in the continuous frames. In this paper, we propose an efficient method to conduct video translation that can preserve the frame modification trends in sequential frames of the original video and smooth the variations between the generated frames. To constrain the consistency of the mentioned tendency between the generated video and the original one, we propose a tendency-invariant loss to impel further exploitation of spatial-temporal information. Experiments show that our method is able to learn more abundant information of adjacent frames and generate more desirable videos than the baselines, i.e., Recycle-GAN and CycleGAN.

Keywords Video translation · Generative adversarial networks · Unsupervised · Spatial-temporal information

1 Introduction

Nowadays, image translation is being widely-used to synthesize vivid pictures into required styles, e.g., we can acquire realistic photographs in styles of Van Gogh and Monet paintings. Likewise, we can also use a similar idea to translate the videos, which contain much more abundant spatial-temporal information compared to static image data. Namely, images are separated from each other, while the frames of a video are tightly correlated. As a result, the methods for static image translation can hardly meet the requirement of video synthesis, as they do not take into consideration the spatial-temporal continuity of the video frames.

There are many approaches that try to synthesize videos into new styles using generative adversarial networks (GANs) [12]. Wang et al. [37] propose a video-to-video synthesis approach using GAN framework as well as the spatial-temporal adversarial objective to synthesize high-resolution and temporally coherent videos, which calls for the input of paired data. As for unpaired videos, Bansal et al. [2] combine spatial–temporal information along with adversarial losses for content translation and style preservation. These methods both exploit the frame continuity in the videos, which manifest better performances than previous method that just utilizes the information in single frames [34]. Nonetheless, they tend to concentrate on maintaining the difference between adjacent frames, while ignoring the fact that the trend of frame modification in the generated video is also supposed to remain the same as in the original video. In other words, these methods are able to preserve some of the features in the source video, yet some information may be lost during the translation. For instance, in the video translation task John Oliver to Stephen Colbert presented in Recycle-GAN, the mouth motions of Oliver are captured, yet some face expressions in the source video are lost after translation. Besides,

✉ Huajun Liu
huajunliu@whu.edu.cn

✉ Dian Lei
dian_lei@whu.edu.cn

Chao Li
ldl1118@whu.edu.cn

Qing Zhu
zhuq66@263.net

[1] School of Computer Science, Wuhan University, Wuhan, China

[2] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China

Published online: 22 July 2020

we also observe that the above video translation methods fail to maintain the smoothness between video frames.

To this end, we are motivated to make use of the consistent trend to synthesize videos that are able to preserve the content information in the source video. Note that the 'content' in our video synthesis method represents the motions or postures in the original video, which ought to remain as much as possible in the generated video. Based on this motivation, we propose an efficient method by involving an additional constraint, the tendency-invariant loss, to ensure the consistency of the trend mentioned above. The experiments show that our method not only synthesizes videos of higher quality and continuity than baselines, but also makes better use of the spatial–temporal information and preserves more content of the source domain. In essence, our contributions are threefold:

- We propose an effective method to make better exploitation of the spatial–temporal information in video frames.
- We put forward a novel loss to improve the photorealism and continuity of the translated videos, with more content of the source video preserved.
- Abundant experiments show that our method is able to learn more abundant spatial-temporal information and perform better than the baselines.

2 Related work

Generative Adversarial Networks (GANs) The vanilla GAN [12] adopts the idea of adversarial learning so that the generated images are indistinguishable from the real ones. Based on this, CGAN [23] is proposed to guide the generation under certain conditions, which has been widely exploited in image generation methods. Later on, many researches dig further into the improvements of conditional GAN [6,8,17,22,25,42].

Conditional GAN can take various forms of input data, such as images [15,20,26,46], categorical labels [6,24,25,43], textual descriptions [27,41,44] and videos [7,11,36,37,45]. Our work, as a method of video translation, also adopts the conditional generative adversarial networks.

Image-to-image translation aims to transfer an image from source domain to target domain, with styles modified and contents preserved. There are many researches exploring this field [4,5,9,14,15,20,21,26,31,32,38,46,47]. Some of these methods require the presence of paired images during training [15], while others alleviate the dependency on paired data and achieve unsupervised translation [9,14,20,46]. UNIT [20] employs variational auto-encoder (VAE) [18] to conduct image-to-image translation under the assumption that images from two different domains can be mapped to the same latent space. CycleGAN [15] first introduces a cycle consistency loss, aiming to learn two mappings such that the translated source samples are inverses to the original samples. MUNIT [14] and DRIT [19] independently propose the assume that images of different domains can be mapped into separated latent spaces, namely the content space and the style space, which can be leveraged to synthesize images belonging to the target domain in a multimodal way. Meanwhile, various approaches are exploring the multi-domain image-to-image translation [1,9], which also take unpaired images as input. In addition, methods like FUNIT [21] pay attention to alleviating the reliance on a large amount of training data and achieving few-shot unsupervised image translation. The idea of cycle consistency is commonly used in unsupervised image-translation methods, which can reduce the arbitrary space of generator in GANs. With respect to video-to-video translation, different methods have resorted to a similar idea to achieve unsupervised translation, so does our method.

Video-to-video translation targets at transferring a video from one style into another, with the content preserved, e.g., human gestures and object motions. Recent work [35] adopts temporal models to make predictions of long-term future poses from a single frame. MoCoGAN [33] improves the results of video generation via decomposing motion and content. Similarly, Temporal GAN [30] employs a temporal generator together with an image generator, in order to generate a set of latent variables and image sequences, respectively. Bashkirova et al. [3] propose a spatio–temporal 3D translator to handle the translation of video implement. Wang et al. [37] propose a video-to-video synthesis approach the spatial–temporal adversarial objective to synthesize high-resolution and temporally coherent videos, which tackles the resolution improvements of the generated videos. Meanwhile, Recycle-GAN [2] focuses on a general video-to-video translation by employing an idea similar to image-to-image translation, where the output is under the control of the input. What's more, Wang et al. [36] achieved few-shot generalization capacity via a weight generation module with an attention mechanism, which explores the few-shot problem and alleviate the reliance of training data. The above approaches are aware of leveraging the spatial-temporal information in videos, yet they focus more on the differences between adjacent frames.

Spatial-temporal constraint for video synthesis Many researches have put emphasis on the spatial–temporal information in the videos [16,39,40]. Kang et al. [16] propose a framework for video object detection, which consists of a tubelet proposal network to generate spatiotemporal proposals, and a long short-term memory (LSTM) network that incorporates temporal information from tubelet proposals to improve the accuracy of object detection in videos. Xiao et al. [40] propose a MatchTrans module to align the spatial–temporal memory from frame to frame. Wang et al. [39]

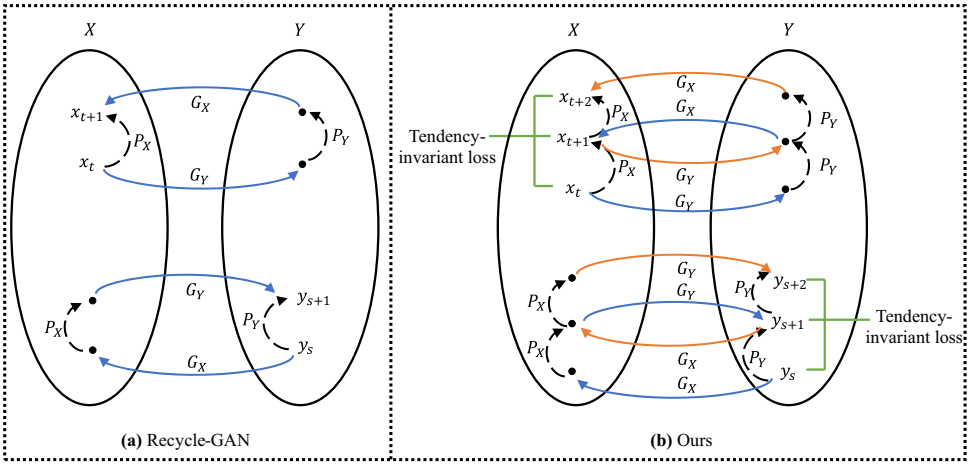


Fig. 1 The key idea of our method compared with Recycle-GAN. **a Recycle-GAN** presents a way to deal with the unpaired but ordered streams $(x_1, x_2, \ldots, x_t, \ldots)$ and $(y_1, y_2, \ldots, y_t, \ldots)$ using spatial–temporal constraint, which is the recycle loss introduced in their work. The whole cycle is across domain and time, and the L1 loss of the frame generated by G_X and the real frame x_{t+1} is expected to be minimized. **b Our method**, based on Recycle-GAN, not only considers the modification between adjacent frames, but also the modification tendency of the generated frames. The newly designed tendency-invariant loss is minimized to ensure the modification trend of source frames is well-preserved in the generated ones

represent videos as space-time region graphs to capture the spatial-temporal information, whose nodes are connected by similarity relations and spatial–temporal relations between objects. These approaches explore the significant spatial–temporal information of videos from different perspectives, mainly by proposing a new framework to capture it.

Our method, as a video synthesis model taking unpaired data as input, refers to the architecture of Recycle-GAN. However, different from Recycle-GAN, we are highly aware of the spatial-temporal continuous information in the video, especially the invariant tendency of the modification between predicted frames and the real ones. The proposed method with the designed loss aims to synthesize videos that are able to preserve the content information in the source video as much as possible.

3 Method

The main contributions of our work are the better exploitation of the spatial–temporal information and the introduction of the tendency-invariant loss. As our model architecture is based on Recycle-GAN, we illustrate the key idea of our work in contrast with it in Fig. 1.

We aim to convert a sequence of source domain images, $X : x_1^T \equiv x_1, x_2, \cdots, x_T$, to a sequence of output images, $\tilde{y}_1^T \equiv \tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_T$ in a way that the style of $\tilde{y}_1^T$ is similar to $Y : y_1^T$, where $x_1^T \in X$, and $y_1^T \in Y$, with the content preserved as much as possible. Thus, the task is to learn a mapping $G_Y : X \rightarrow Y$. Note that our model takes unpaired video frames as input during training.

3.1 Full object

Since we adopt the architecture of GAN, the vanilla adversarial loss is used as well, termed as $\mathcal{L}_{\text{GAN}}$ in our work. And the cycle consistency loss $\mathcal{L}_{\text{cycle}}$ in CycleGAN [46] is adopted. Besides, the recurrent loss $\mathcal{L}_{\text{recurrent}}$ and the recycle loss $\mathcal{L}_{recycle}$ in Recycle-GAN [2] are also leveraged. Meanwhile, we introduce a novel tendency-invariant loss $\mathcal{L}_{\text{inv}}$ to impel the model and improve the whole translation. The full loss function of our work is as follows:

$$\begin{aligned}\min_{G,P}\max_{D}\mathcal{L}_{\text{all}}(G,P,D) &= \mathcal{L}_{\text{GAN}}(G_X,D_X)\\ &+\mathcal{L}_{\text{GAN}}(G_Y,D_Y)\\ &+\alpha\mathcal{L}_{\text{cycle}}(G_X,G_Y)+\alpha\mathcal{L}_{\text{cycle}}(G_Y,G_X)\\ &+\beta\mathcal{L}_{\text{recurrent}}(P_X)+\beta\mathcal{L}_{\text{recyrrent}}(P_Y)\\ &+\gamma\mathcal{L}_{\text{recycle}}(G_X,G_Y,P_X)\\ &+\gamma\mathcal{L}_{\text{recycle}}(G_X,G_Y,P_Y)\\ &+\delta\mathcal{L}_{\text{inv}}(P_X)+\delta\mathcal{L}_{\text{inv}}(P_Y)\end{aligned} \tag{1}$$

where the parameters α, β, γ and δ are used to balance the function. It is worth noting that the tendency-invariant loss $\mathcal{L}_{\text{inv}}$ is the key insight of our method, which significantly improves the exploitation of the spatial–temporal information in video frames. Next, these five components would be illustrated in detail, and we only show one side mapping of the cycle process for clarity.

3.2 Adversarial loss

In the process of adversarial training, we train a discriminator D_Y to distinguish the generated sample $G_Y(x_t)$ from a real sample y_s. Meanwhile, the generator G_Y is expected to generate results that can fool D_Y. The minimax game can be expressed as:

$$\min_{G_Y}\max_{D_Y}\mathcal{L}_{\text{GAN}}(G_Y, D_Y) = \sum_s \log D_Y(y_s) + \sum_t \log(1 - D_Y(G_Y(x_t))), \tag{2}$$

3.3 Cycle consistency loss

We only use unpaired samples individually in respective videos during training, without the requirement of paired input data. To tackle this, the cycle consistency loss is essential and leveraged by our method, which can be written as:

$$\min_{G_X,G_Y}\mathcal{L}_{\text{cycle}}(G_X, G_Y) = \sum_t \|x_t - G_X(G_Y(x_t))\|_1 \tag{3}$$

3.4 Recurrent loss

The above two losses are identical to the losses for static image translation. In order to deal with video data, we must take advantage of the temporal ordering of the sequential frames. We adopt a recurrent temporal predictor P_X introduced in Recycle-GAN [2], which is supposed to give prediction of future frames given the past frame information. The recurrent loss is expressed as:

$$\min_{P_X}\mathcal{L}_{\text{recurrent}}(P_X) = \sum_t \|x_{t+1} - P_X(x^t_{t-1})\|_1 \tag{4}$$

where $P_X(x^t_{t-1})$ represents the prediction of P_X given x_{t-1} and x_t as the input.

3.5 Recycle loss

Combining the image generator and the temporal prediction network, the recycle loss across domains and time can be written as:

$$\min_{G_X,G_Y,P_Y}\mathcal{L}_{recycle}(G_X, G_Y, P_Y) = \sum_t \|x_{t+1} - G_X(P_Y(G_Y(x^t_{t-1})))\|_1 \tag{5}$$

where $G_Y(x^t_{t-1})$ denotes $G_Y(x_{t-1})$ and $G_Y(x_t)$. P_Y is fed with the output of G_Y, then forwards its prediction to G_X. The whole cross-domain and cross-time cycle is referred to as a recycle round, and the frame generated by G_X would be compared with the real frame x_{t+1} and evaluated the L1 loss. The recycle loss is the essence of Recycle-GAN, which exploits the temporal information of video frames.

3.6 Tendency-invariant loss

As an improvement of Recyle-GAN, we dig further into the temporal information of the video. Apart from the above losses, we come up with a tendency-invariant loss, under the assumption that the frame modification trend ought to be maintained in the predicted frames with the aim to preserve the content of source domain and improve the quality of the output. According to this, our tendency-invariant loss is designed as follows:

$$\min_{P_X}\mathcal{L}_{\text{inv}}(P_X) = \sum_t \| \left(\left(P_X\left(x^{t+1}_t\right) - P_X\left(x^t_{t-1}\right)\right) - \left(P_X\left(x^t_{t-1}\right) - x_t\right)\right) - \left(\left(x_{t+2} - x_{t+1}\right) - \left(x_{t+1} - x_t\right)\right) \|_1 \tag{6}$$

where $P_X(x^{t+1}_t)$ denotes the prediction of frame x_{t+2} given x_t and x_{t+1} as input, and $P_X(x^t_{t-1})$ denotes the prediction of x_{t+1} given x_{t-1} and x_t. By calculating the difference of two L1 losses, we intend to obtain the tendency of the frame modification, instead of merely the modification itself. Minimizing the tendency-invariant loss means that we expect the frame modification tendency can be preserved in the predicted frames.

4 Experiments

In this section, we will introduce the experimental setups, implementation details and the experimental results of our work.

4.1 Experimental setups

We illustrate our experimental setups including datasets, baselines and the evaluation metrics in this part.

4.1.1 Datasets

Although the Cityscapes dataset [10] is commonly used in other work, we consider it improper to be used in our experiments. For one thing, the labeled images in Cityscapes are not in the form of continuous sequences. For another, Recycle-GAN uses the Viper dataset to evaluate its effectiveness as well. Hence, we choose Viper dataset and the other two datasets that Recycle-GAN provides to conduct our experiments and make the evaluations.

– Viper [28] dataset is a publicly available dataset collected by computer games with realistic scene and fine-annotated pixel-level labels, which has also been used in Recycle-GAN. We use it to evaluate the methods for the task images ↔ labels. The vision of Grand Theft Auto V game, which simulates a functioning city, is adopted for our experiments. Please note that the purpose of demonstrating this task is not to achieve high segmentation accuracy, since the baselines and ours are not proposed for video segmentation. Instead, we make this comparison solely for a better understanding of our introduced tendency-invariant loss. There are 77 different video sequences containing 5 diverse scenes, i.e., *day, sunset, rain, snow, night*. From those, we use 57 sequences for training and 20 sequences for evaluation. The original resolution of the images is 1920 × 1080, which is then resized to 256 × 256 to be fed into the model during training and testing.
– Face-to-Face dataset and Flower-to-Flower dataset, both of which are made public by Recycle-GAN, are also used in our experiments. The Face-to-Face dataset is composed of various videos of celebrity faces, which are acquired via the facial keypoints extracted by the OpenPose Library. We use this dataset to accomplish the video translation tasks, Barack Obama ↔ Donald Trump and John Oliver ↔ Stephen Colbert. The Flower-to-Flower dataset contains the time-lapse of various blooming flower from public videos. There are 4 different sequences in this dataset, each of which contains about 200 frames, and the flower blooming orders are not synchronous.

4.1.2 Baselines

To verify the effectiveness of our model, we compare it with models that focus on video translation with GANs. Since our model architecture is based on Recycle-GAN and take unpaired video data as input, we choose CycleGAN [46] and Recycle-GAN [2] as the baselines of our experiments.

– CycleGAN [46] uses two generators to translate images, with the idea of cycle consistency. We employ it to translate the video frames and make comparisons, with the intent to figure out the effect of the spatio–temporal constraint.
– Recycle-GAN [2] uses two generators and two predictors to translate videos. It puts forward a recycle loss to work with cycle loss and recurrent loss for content translation and style preservation, which takes the temporal information into consideration. The contrasts with Recycle-GAN aim to demonstrate the significant improvements that our method makes about the spatial–temporal information.

4.1.3 Evaluation metrics

Different aspects of experiments have been implemented to show the superiority of our method. Hence, we would discuss the metrics according to the respective fields. Note that all the scores of the metric results are calculated in mean values.
Segmentation accuracy. As mentioned before, the Viper dataset contains groundtruth images; thus, we can evaluate the segmentation performances of different methods. The following three metrics are commonly used for image segmentation task.

– Pixel Accuracy (PA): it calculates the ratio between the amount of properly classified pixels and the total number of them.
– Mean Pixel Accuracy (MPA): it is the transform of PA, in which the ratio of correct pixels is computed in a per-class basis and then averaged over the total number of classes.
– Mean Intersection over Union (mIoU): it is the standard metric for image segmentation. It computes a ratio between the intersection and the union of two sets, i.e., the synthesized label videos and the ground-truth label videos.

Normalized FCN score A pre-trained FCN-style model is employed to evaluate the similarity between generated images and real ones, which has been used in Recycle-GAN as well. Higher scores indicate higher quality of an approach that can synthesize samples more similar to real data.
Fréchet Inception Distance (FID) This metric is based on Fréchet Inception Distance (FID) [13], which aims to measure the similarities between two sets of images. It evaluates whether the generated samples are well translated. And a lower FID score represents a translation result with better diversity and realistic. In our experiment, it indicates the quality of the video translation.
User Study We perform a user study on the synthesized outputs to evaluate the effectiveness of our method. Twenty-seven participants are selected to fulfill the study. They are shown the synthesized videos of Recycle-GAN and ours simultaneously and are asked which one has higher quality, better smoothness and better stability between video sequences.

4.2 Implementation details

We design our model based on the structure of Recycle-GAN [2], where the generators are the same as CycleGAN [46] and the predictors are the same as Pix2Pix [15]. The learning rate of our approach is set as 0.0002, and the tendency-invariant loss is calculated using 4 consecutive frames, and the batch size is 1. There are two downsampling convolution (each with

Fig. 2 The translation result of Recycle-GAN and our method on Donald Trump → Barack Obama

Fig. 3 The translation result of Recycle-GAN and our method on Stephen Colbert → John Oliver

stride-2), six residual blocks, and two upsampling convolution (each with a stride 0.5) in the G_X and G_Y. The resolution of input images is 256×256 pixels. We concatenate the two frames as the input of P_X and P_Y, for which we adopt the U-Net architecture [15,29]. We set the parameters α, β, γ, δ as 10. The structure of discriminator network is a 70×70 PatchGAN [15]. The three models are trained 15 epochs on Viper (about 62 h), 200 epochs on Flower-to-Flower (about 8 h), and 60 epochs on Face-to-Face (Obama ↔ Trump: about 44 h; Oliver ↔ Colbert: about 91 h). Similar to Recycle-GAN, the time of generating one frame by our trained model is 52 ms. All the experiments are executed on a single NVIDIA RTX 2080Ti GPU.

4.3 Experimental results

In this part, we would demonstrate the qualitative and quantitative results of the comparisons against the baselines, for which the analysis is made according to different datasets.

Analysis on Face-to-Face For this dataset, we evaluate on two tasks, i.e., Barack Obama ↔ Donald Trump and John Oliver ↔ Stephen Colbert. We show one direction for task Donald Trump → Barack Obama and both directions for John Oliver ↔ Stephen Colbert. Moreover, please note that the illustration is quite limited in images, we have put the generated videos in our supplementary material and highly suggest readers to refer to the videos for more details.

Fig. 4 The translation result of Recycle-GAN and our method on John Oliver → Stephen Colbert

Table 1 The FID scores of different methods on Obama ↔ Trump and Oliver ↔ Colbert

Model	Obama ↔ Trump	Oliver ↔ Colbert
CycleGAN	56.31	72.14
Recycle-GAN	49.47	53.38
Ours	**25.89**	**49.47**

Bold values indicate the best results in the experiments, in order to emphasize

Figure 2 shows an example of Donald Trump → Barack Obama. Both of the two methods are able to capture the stylistic face expressions of Donald Trump. However, by observing the videos in the supplementary material, we should notice that the video generated by Recycle-GAN is not stable compared to ours. What's more, the synthesized frames of our output are of higher quality and better photorealism.

Figure 3 shows the results of Stephen Colbert → John Oliver; our method has better performances than Recycle-GAN in frame quality and video stability, and the face expressions are better captured by ours as well.

On John Oliver → Stephen Colbert shown in Fig. 4, the detailed expressions are more correctly learned by our method than Recycle-GAN. For example, in Recycle-GAN, the motion trends of eyes and mouth vary obviously from the corresponding original frames, while ours preserves the source content quite well.

Besides, our method has slight advantages over the baselines on FID scores, as shown in Table 1. For instance, on Stephen Colbert → John Oliver, the glasses in Recycle-GAN's output video are not well-translated.

Analysis on Viper As mentioned before, the task of images ↔ labels is demonstrated to show the effect our exploitation of modification tendency; thus, we would concentrate more on the improvements of the video output shown in Fig. 5, and the metric results of semantic segmentation accuracy in Table 2 are just for reference.

On images → labels, we can notice that the differences between adjacent frames are quite apparent in the video generated by Recycle-GAN, e.g., the semantic classification of the road is not stable along the sequences, which should not be assigned as brown. In contrast, the frame sequences of our method perform a much more continuous video, as every two adjacent frames tend to maintain the changing tendencies in the source video. As for the scores on the commonly used metrics for semantic segmentation, we can also tell that our method surpasses CycleGAN as well as Recycle-GAN in Pixel Accuracy, Mean Pixel Accuracy and mean Intersection over Union, as shown in Table 2.

Likewise, on labels → images, the cloth color of the man on the motorbike is better learned by our method than Recycle-GAN. In addition, we compare the FCN results of all the methods on different scenes in the Viper dataset, which is shown in Table 3. We can conclude that videos generated by our method manifest the highest similarity to the input videos on 5 different scenes.

Analysis on Flower-to-Flower Figure 6 shows the synthesized frames of our method on Flower-to-Flower dataset. The videos in this dataset depict the blooming of various flowers, which is a relatively slow process, which means that the modifications between adjacent frames are quite slight. Hence, the exploitation of the spatial-temporal information by Recycle-GAN is enough to handle the video translation, so the outputs of Recycle-GAN and ours do not differ much from each other.

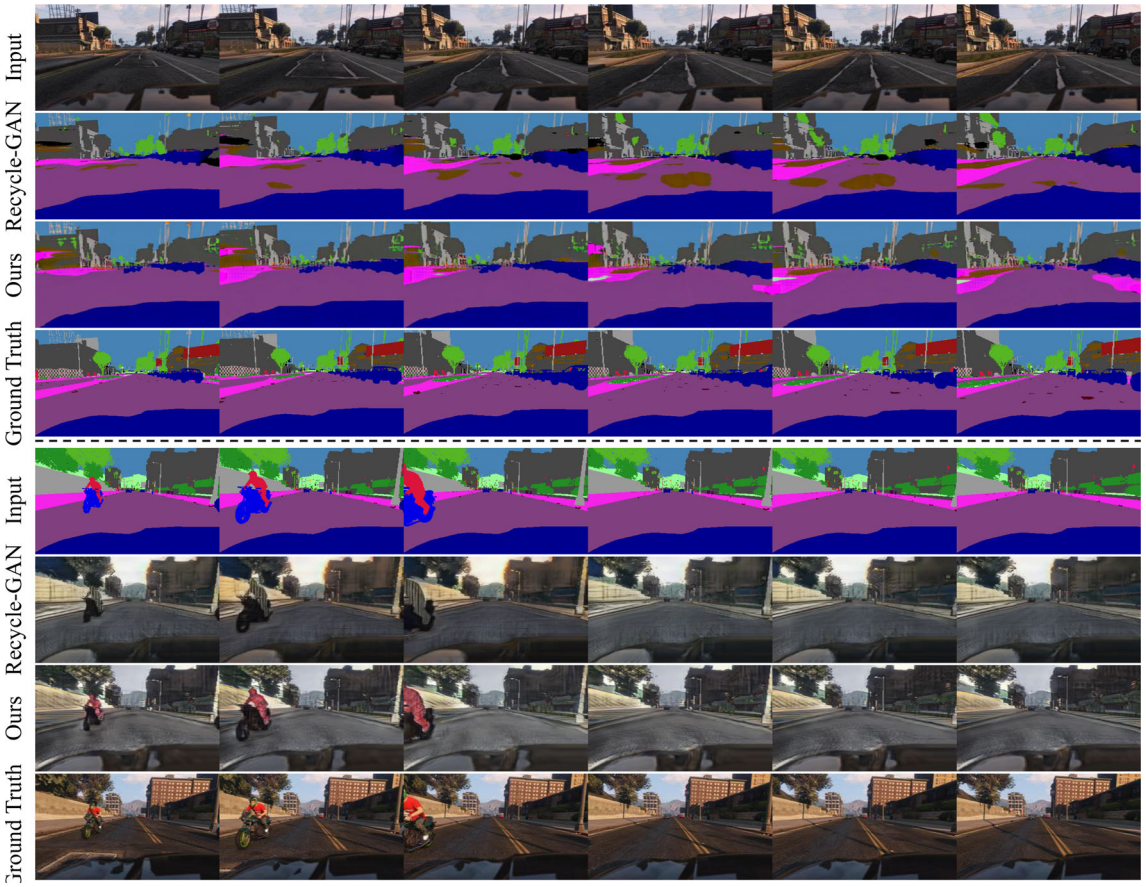


Fig. 5 The video translation results of Recycle-GAN and our method on Viper dataset

Table 2 The semantic segmentation metrics of different methods on Viper dataset

Criterion	Model	Day	Sunset	Rain	Snow	Night	All
Pixel accuracy	Cycle-GAN	0.362	0.412	0.439	0.296	0.223	0.349
	Recycle-GAN	0.442	0.572	0.591	0.560	0.469	0.503
	Ours	**0.497**	**0.616**	**0.599**	**0.624**	**0.510**	**0.548**
Mean pixel accuracy	Cycle-GAN	0.069	0.071	0.046	0.087	0.051	0.065
	Recycle-GAN	0.122	0.151	0.099	0.136	**0.073**	0.116
	Ours	**0.153**	**0.163**	**0.129**	**0.140**	**0.073**	**0.135**
mIoU	Cycle-GAN	0.053	0.042	0.069	0.043	0.019	0.047
	Recycle-GAN	0.076	0.101	0.068	0.083	0.045	0.074
	Ours	**0.083**	**0.129**	**0.078**	**0.091**	**0.067**	**0.087**

Bold values indicate the best results in the experiments, in order to emphasize

Table 3 The normalized FCN scores of different methods on Viper dataset

Model	Day	Sunset	Rain	Snow	Night	All
CycleGAN	0.36	0.25	0.28	0.31	0.33	0.32
Recycle-GAN	0.37	0.37	**0.39**	0.36	**0.42**	0.38
Ours	**0.39**	**0.41**	**0.39**	**0.39**	0.41	**0.40**

Bold values indicate the best results in the experiments, in order to emphasize

As for the FID scores shown in Table 4, we manifest slight advantages over Recycle-GAN, due to the improvements of video continuity and stability brought by the spatial-temporal constraint.

Analysis on user study The results of user study are shown in Table 5. The scores tell that a majority of the participants prefer the our synthesized videos than that of Recycle-GAN, which proves our outputs are better in smoothness due to the tendency-invariant loss.

Fig. 6 The translation result of Recycle-GAN and our method on Flower-to-Flower dataset

Table 4 The FID scores of different methods on Flower-to-Flower dataset

Model	Flower 1	Flower 2	Flower 3	Flower 4	All
CycleGAN	123.5	133.1	160.2	99.6	129.1
Recycle-GAN	117.3	120.7	139.0	96.2	118.3
Ours	**112.6**	**116.9**	**129.9**	**89.3**	**112.2**

Bold values indicate the best results in the experiments, in order to emphasize

Table 5 The results of user study on different datasets

User study	Viper		Obama ↔ Trump		Oliver ↔ Colbert		Flowers	
	Recycle	Ours	Recycle	Ours	Recycle	Ours	Recycle	Ours
	43.6	**56.4**	30.3	**69.7**	39.6	**60.4**	47.4	**52.6**

Bold values indicate the best results in the experiments, in order to emphasize

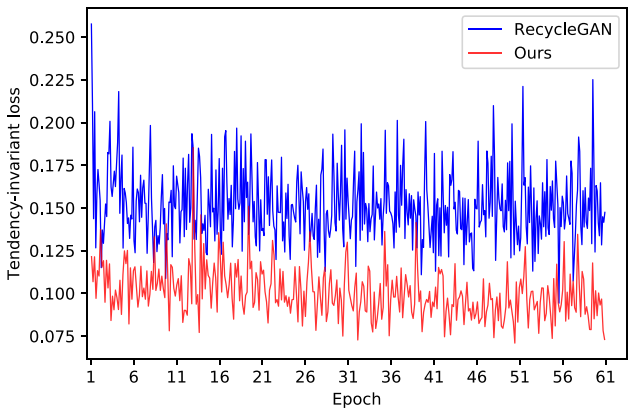


Fig. 7 The comparison of the tendency-invariant loss during training

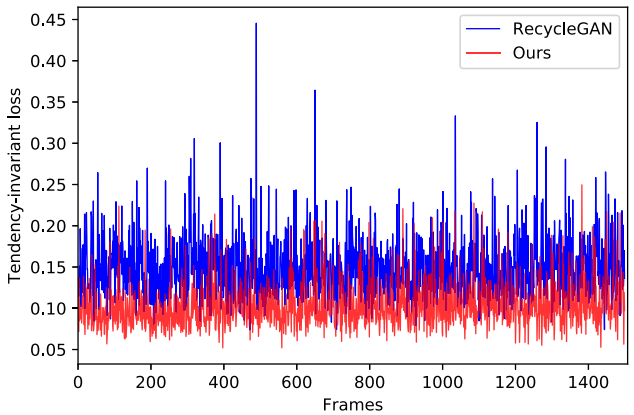


Fig. 8 The frame-by-frame comparison of the tendency-invariant loss on one tested video sequence

Analysis on tendency deviation To evaluate the tendency deviation of our proposed method, we train our model and Recycle-GAN on Barack Obama ↔ Donald Trump and record the tendency-invariant loss calculated by Equation 6. In Fig. 7, the loss trend during training is showed, and we can see that the tendency-invariant loss of ours is much lower than Recycle-GAN's. Figure 8 shows the frame-by-frame comparison on one tested video sequence, where ours is also lower than Recycle-GAN. These comparisons are meant to demonstrate that our proposed loss is able to help maintain the modification tendency between frames and achieve better smoothness.

5 Limitation

Our method is proposed to tackle the issues of content preservation and video continuity during translation; thus, we focus more on the modification trend between frames. For datasets that depict tardy and slight motions of subjects, the superiority of our spatial-temporal constraint may not be apparent compared to Recycle-GAN. This is the inherent limitation of our method, which we intend to work on in the future.

6 Conclusions

In this paper, we propose an effective method to tackle the video synthesis using GANs, based on the architecture of Recycle-GAN. However, we are highly aware of the spatial–temporal continuous information in the video, especially the invariant tendency of the modification between predicted frames and the real ones. A novel tendency-invariant loss is designed to constrain our model to synthesize videos that are more stable and preserve more content in the source video.

Acknowledgements This work was funded by the National Natural Science Foundation of China (NSFC) (41771427 and 41631174).

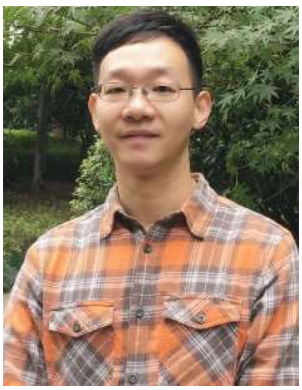
Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Anoosheh, A., Agustsson, E., Timofte, R., Van Gool, L.: Combogan: unrestrained scalability for image domain translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 783–790 (2018)
2. Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: unsupervised video retargeting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 119–135 (2018)
3. Bashkirova, D., Usman, B., Saenko, K.: Unsupervised video-to-video translation. arXiv:1806.03698 (2018)
4. Benaim, S., Wolf, L.: One-shot unsupervised cross domain translation. In: Advances in Neural Information Processing Systems, pp. 2104–2114 (2018)
5. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3722–3731 (2017)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019)
7. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5933–5942 (2019)
8. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2172–2180 (2016)
9. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
11. Gafni, O., Wolf, L., Taigman, Y.: Vid2game: controllable characters extracted from real-world videos. arXiv:1904.08379 (2019)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
14. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
16. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 727–735 (2017)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
19. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018)
20. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
21. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. arXiv:1905.01723 (2019)

22. Ma, T., Tian, W.: Back-projection-based progressive growing generative adversarial network for single image super-resolution. Vis. Comput (2020). https://doi.org/10.1007/s00371-020-01843-3
23. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014)
24. Miyato, T., Koyama, M.: cgans with projection discriminator. arXiv:1802.05637 (2018)
25. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, pp. 2642–2651. JMLR.org (2017)
26. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)
27. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv:1605.05396 (2016)
28. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2213–2222 (2017)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
30. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2830–2839 (2017)
31. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2107–2116 (2017)
32. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv:1611.02200 (2016)
33. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535 (2018)
34. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: forecasting from static images using variational autoencoders. In: European Conference on Computer Vision, pp. 835–851. Springer (2016)
35. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: video forecasting by generating pose futures. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3332–3341 (2017)
36. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. arXiv:1910.12713 (2019)
37. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv:1808.06601 (2018)
38. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
39. Wang, X., Gupta, A.: Videos as space-time region graphs. In: The European Conference on Computer Vision (ECCV) (2018)
40. Xiao, F., Jae Lee, Y.: Video object detection with an aligned spatial-temporal memory. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 485–501 (2018)
41. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)
42. Yuan, Q., Li, J., Zhang, L., Wu, Z., Liu, G.: Blind motion deblurring with cycle generative adversarial networks. Vis. Comput. **36**, 1591–1601 (2019)
43. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv:1805.08318 (2018)
44. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915 (2017)
45. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Dance dance generation: motion transfer for internet videos. arXiv:1904.00129 (2019)
46. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
47. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, pp. 465–476 (2017)



Huajun Liu received Ph.D. degree from School of Computer at Wuhan University. Now he is an associate professor in School of Computer Science, Wuhan University, P.R.China. In recent years, Prof. Liu's research interests include virtual geographic environments, computer vision, computational photography and deep learning.

Chao Li received B.S. degree from Lanzhou University. Now he is pursuing M.S. degree in School of Computer Science, Wuhan University, P.R.China. His research interests include computer vision, deep learning and video translation.

Dian Lei received B.S. degree from Nanjing University of Aeronautics and Astronautics. Now she is pursuing M.S. degree in School of Computer Science, Wuhan University, P.R.China. Her research interests include computer vision, deep learning and image processing.

Qing Zhu is Chang Jiang Scholars professor in photogrammetry and GIS, Professor Committee Director, Faculty of Geosciences and Environmental Engineering of Southwest Jiaotong University, P.R.China. In recent years, Prof. Zhu's research interests include digital terrain modeling (DEM), three-dimensional geographic information system (3D GIS) and virtual geographic environments (VGE).