Video-to-Video Translation with Global Temporal Consistency

Xingxing Wei
Department of Computer Science and Technology
Institute for Artificial Intelligence
State Key Laboratory for Intelligent Technology and Systems
Tsinghua Lab of Brain and Intelligence
Tsinghua University
Haidian, Beijing, China
xwei11@mail.tsinghua.edu.cn

Jun Zhu *
Department of Computer Science and Technology
Institute for Artificial Intelligence
State Key Laboratory for Intelligent Technology and Systems
Tsinghua Lab of Brain and Intelligence
Tsinghua University
Haidian, Beijing, China
dcszj@mail.tsinghua.edu.cn

Sitong Feng
Faculty of Information Technology
Macau University of Science and Technology
Macau, China
sitongfeng@gmail.com

Hang Su
Department of Computer Science and Technology
Institute for Artificial Intelligence
State Key Laboratory for Intelligent Technology and Systems
Tsinghua Lab of Brain and Intelligence
Tsinghua University
Haidian, Beijing, China
suhangss@mail.tsinghua.edu.cn

ABSTRACT

Although image-to-image translation has been widely studied, the video-to-video translation is rarely mentioned. In this paper, we propose an unified video-to-video translation framework to accomplish different tasks, like video super-resolution, video colourization, and video segmentation, etc. A consequent question within video-to-video translation lies in the flickering appearance along with the varying frames. To overcome this issue, a usual method is to incorporate the temporal loss between adjacent frames in the optimization, which is a kind of *local* frame-wise temporal consistency. We instead present a residual error based mechanism to ensure the video-level consistency of the same location in different frames (called *global* temporal consistency). The global and local consistency are simultaneously integrated into our video-to-video framework to achieve more stable videos. Our method is based on the GAN framework, where we present a two-channel discriminator. One channel is to encode the video RGB space, and another is to encode the residual error of the video as a whole to meet the global consistency. Extensive experiments conducted on different video-to-video translation tasks verify the effectiveness and flexibleness of the proposed method.

KEYWORDS

Video-to-Video Translation, Temporal Consistency, Generative Adversarial Network

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240708

ACM Reference Format:
Xingxing Wei, Jun Zhu [1], Sitong Feng, and Hang Su. 2018. Video-to-Video Translation with Global Temporal Consistency. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240708

1 INTRODUCTION

Recently, the so-called image-to-image translation has been widely studied [9, 24], i.e., an image is input to a generative network, and the output is also an image. Many tasks can be formulated as the image-to-image translation problem, such as image super-resolution [21], image colourization [23], image segmentation [4], and so on. Up to now, the Generative Adversarial networks (GAN) methods are dominantly used to solve this problem. Although much progress has been achieved in image-to-image translation, the video-to-video translation is rarely mentioned. In fact, the above super-resolution, stylization, and segmentation tasks are easily extended to the video case, result in the video super-resolution [20], video colourization [1], and video segmentation [10] , respectively. Therefore, like the image-to-image translation, an unified video-to-video translation framework should be considered to handle with these different tasks.

To accomplish the video-to-video translation, a straightforward method is to perform image-to-image translation on each frame in the video. However, this kind of operation will lead to the flickering results, which is usually stated in the video stylization [2]. Specifically, the same location within the different frames will show slightly difference on the stylization appearance. As a result, the generative video is flickering along with the varying frames. We argue that this phenomenon not only occurs in the video stylization, but also all the video-to-video translation tasks, including the video segmentation, video super-resolution, and video colourization (see the rectangles in Fig.1). How to overcome this problem and obtain stable videos becomes a major issue to be addressed in video-to-video translation.

*Corresponding author.

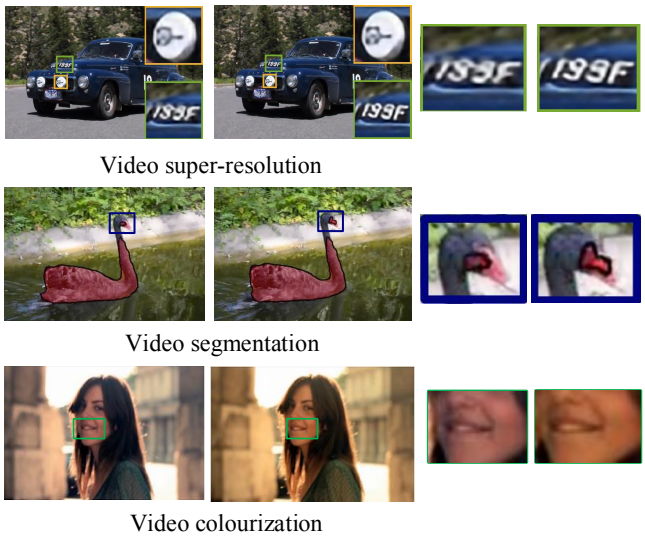

Figure 1: Two consecutive frames from the results of video super-resolution [20], video segmentation [10] and video colourization with per-frame processing [1], respectively. From the figure, we see that the appearance of the same location in consecutive frames is saltatorial and not smoothly changing. This will lead to the flickering results in videos.

In video stylization, the main solution for the unstable videos is to add the temporal consistency loss in the optimization [2, 7, 18]. When the loss is computed between two adjacent frames, it is called short-term consistency, if more adjacent frames are involved, it is called long-term consistency. In the view of technique, because temporal consistency loss ensures the frame-by-frame consistency, it is a kind of *local* frame-wise mechanism, and thus cannot get the video-level consistency. In such situation, we need to evaluate the temporal consistency of the video as a whole, which can be viewed as a *global* mechanism. Combining the local and global consistency together is a more suitable mechanism to obtain the stable videos.

Considering these factors, in this paper, we propose an unified video-to-video translation framework with an integrating global and local mechanism to ensure the consistency of the same location within different frames. Inspired by the image-to-image translation, we also utilize the generative adversarial networks. However, to deal with the video data, the generator and discriminator are both RNN-based architectures (please see section 3.2 for the details). For the inconsistency between frames, we exploit the optical flow to warp the neighboring frames, and compute the residual errors between the warped frames and the aligned frames. The computed residual errors are used in two places to ensure the consistency. The first one is put as a part of the generator's loss function to enforce the residual error's minimization during the training phrase (local temporal consistency), and the second one is fed to the discriminator to guide the generation of video frame with lowest residual errors (global temporal consistency). Different from the traditional GAN, we here present a two-channel discriminator, where one channel is to encode the video RGB space as usual, and another is to encode the residual errors between adjacent frames. During the optimization, the discriminator will compare the predicted videos with the ground-truth videos as a whole, and thus guide the predicted video to be as stable as the ground-truth video. Therefore, the discriminator with the channel of residual errors plays the role of a global guide for the temporal consistency. In this way, the local and global temporal consistency are simultaneously integrated into the video-to-video framework. Fig.2 illustrates the whole architecture of the proposed method.

In summary, this paper has the following contributions:

- To our knowledge, we are the first one to propose an unified framework to accomplish different video-to-video translation problems. To verify the effectiveness, we utilize the proposed framework to perform three tasks: video super-resolution, video colourization, and video segmentation. The proposed framework is sufficiently flexible, and can directly use the existing video generating methods as the generator in our framework.
- We give an analysis about the temporal inconsistency existing in the video-to-video translation tasks, and propose a novel two-channel discriminator to ensure the global temporal consistency for the testing video as a whole. Furthermore, by incorporating the existing local temporal consistency with the proposed global temporal consistency in an unified framework, our method shows a major improvement over the single local consistency (see the experiments).

The rest of this paper is organized as follows. In Section 2, we briefly review the related works. We present the video-to-video translation framework in Section 3. Section 4 reports all experimental results of different video-to-video translation tasks. Finally, we summarize the conclusions in Section 5.

2 RELATED WORK

The related work comes from two aspects: image-to-image translation and temporal consistency.

2.1 Image-to-Image Translation

In computer vision, many tasks can be formulated as the image-to-image translation problem, such as image super-resolution, image segmentation, image denoise, deblure, dehaze, and so on. In the past, these tasks are studied individually. Therefore, although a lot of specific frameworks are proposed [3, 5, 16], they are not compatible with each other. With the rise of deep learning, an universal image-to-image translation framework (called pix2pix) based on the GAN (Generate Adversarial Network) is proposed. The pix2pix method [9] is the first to performs the "Labels to Street Scene", "Aerial to Map", "Day to Night", and "BW to Color" via an unified framework, and achieves a good performance.

However, the pix2pix method needs a lot of image pairs in the training phrase, which usually are not easy to obtain in the real world. This motivates the generation of unpaired image-to-image translation framework. CycleGAN [24] is presented for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. It achieves their goal via learning a cyclic mapping and minimizing the so-called cycle consistency loss. The similar ideas are used in DiscoGAN [12] and Dual GAN [22].

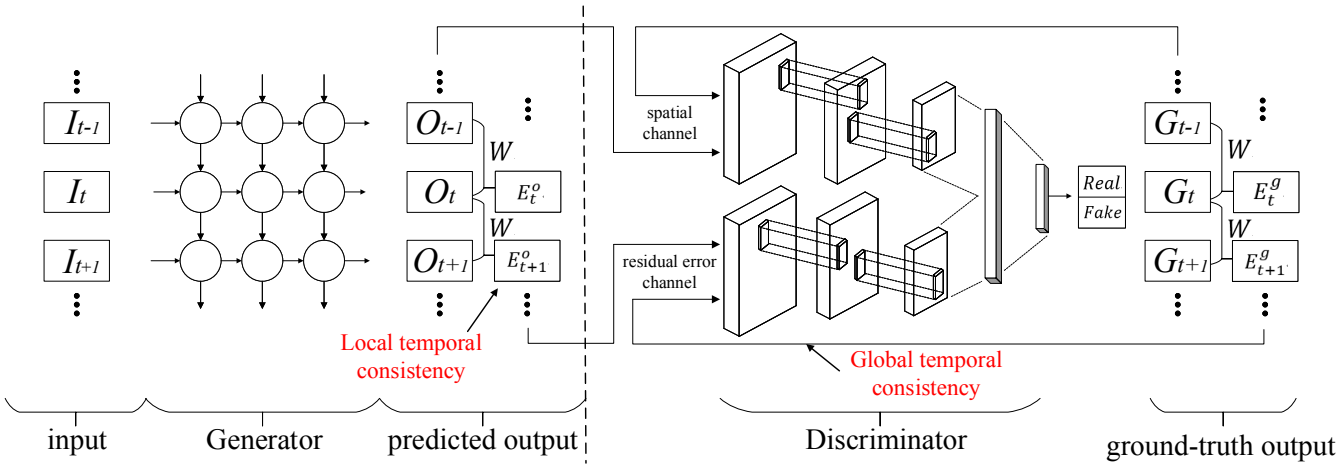

Figure 2: Architecture of the proposed video-to-video translation framework. Our method is based on the GAN. In the generator, I_{t-1}, I_t, I_{t+1} are the input video frames, and O_{t-1}, O_t, O_{t+1} are the generated output video frames. We first compute the optical flow W between the neighboring frames (e.g. I_{t-1}, I_t). Then O_{t-1} is warped using W to produce $\hat{O}_t$ to align O_t. Next, the residual error $E^o_t(E^o_{t+1})$ between the warped frame $\hat{O}_t(\hat{O}_{t+1})$ and aligned frame $O_t(O_{t+1})$ is computed. For the real data (ground truth), we also compute the residual error (e.g. E^g_t, E^g_{t+1}) using the same method. In the two-channel discriminator, we use one channel to encode the ground-truth video frames $\mathcal{G} = \{..., G_{t-1}, G_t, G_{t+1}, ...\}$ and the generated frames $\mathcal{O} = \{..., O_{t-1}, O_t, O_{t+1}, ...\}$. We use another channel to encode the residual errors from predicted output $\mathcal{E}^g = \{..., E^g_t, E^g_{t+1}, ...\}$ and the ground-truth output $\mathcal{E}^o = \{..., E^o_t, E^o_{t+1}, ...\}$. The outputs of these two channels are finally concatenated to give the discriminator's result.

Neural Style Transfer (NST) [6, 11, 14] is another way to perform image-to-image translation, which synthesizes a novel image by combining the content of one image with the style of another image (typically a painting) based on matching the Gram matrix statistics of pre-trained deep features. A major difference between NST and pix2pix method is that NST has no ground-truth targeted images in the training. Therefore, the generated results cannot be evaluated quantitatively like pix2pix method.

Unlike the above work, we focus on the video-to-video translation problem, which is an extension of the image-to-image translation. Similarly, many computer vision tasks can formulated into such a framework, such as video segmentation, video super-resolution, and video colourization, and so on. Therefore it is worthy to be further studied.

2.2 Temporal Consistency

Neural Style Transfer (NST) demonstrates that deep learning can produces fantastic stylized images with the appearance of a given artwork. It is straightforward to extend the idea to videos. Actually many researchers [2, 7, 18] have explored this task. The experiments show that simply applying image NST techniques to video frames is non-trivial, and often leads to flickering results. Therefore, a well-designed method is needed to ensure the temporal consistency.

[7] and [18] use a short-term and long-term temporal loss to solve this issue. Specifically, they utilize the optical flow to warp the previous frame to align the current frame (short-term), and then compute the loss between the warped frame and current frame. By minimizing the temporal loss, one can get the stable video stylization. If more distant previous frames are used, a long-term loss is obtained. Note that the temporal loss is computed directly on the stylized frames. Instead, [2] computes the temporal loss on the feature maps output by the encode network, and propose an efficient network by incorporating short-term coherence, and propagating short-term coherence to long-term, which ensures consistency over a longer period of time. Prior to this, [1] explores the temporal consistency in video enhancement and depth estimation. They propose a gradient-domain technique to generate a temporally-consistent video sequence. The core of their solution is to infer the temporal regularity from the original unprocessed video, and use it as a temporal consistency guide to stabilize the processed sequence.

Essentially, both the temporal loss and temporal regularity are a kind of local information. They achieve the video-level consistency by ensuring the inter-frames consistency. As a contrast, we use not only the local constraint, but also the discriminate network in GAN to give a global stable evaluation for the entire video, which is more comprehensive.

3 THE PROPOSED FRAMEWORK

In this section, we give the details of the proposed framework for video-to-video translation with global and local temporal consistency.

3.1 Residual Error based on Optical Flow

As mentioned before, because the generator and discriminator in our framework both use the residual error to enforce the local and global temporal consistency, respectively, we first give the introduction about the residual error based on optical flow.

Given the input video sequence $\{I_t | t = 1...n\}$, the task is to translate $\{I_t | t = 1...n\}$ to the predicted video sequence $\{O_t | t = 1...n\}$,

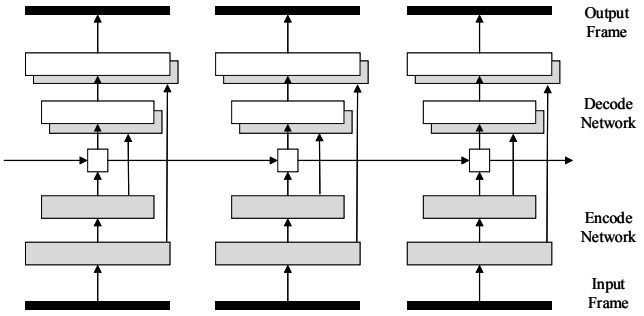

Figure 3: Illustrations of the Generator in our framework. We combine the RNN and U-net [17] to construct our Generator network. For the details, please see the texts.

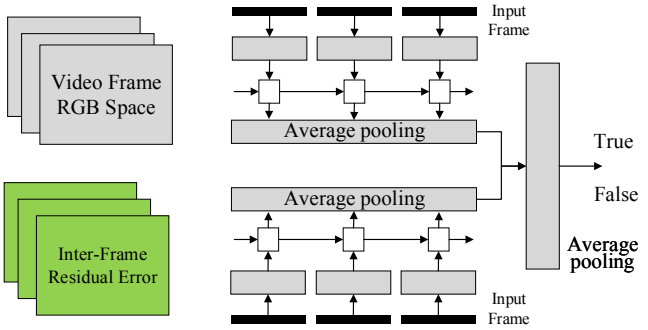

Figure 4: Illustrations of the Discriminator in our framework. We combine the RNN and Markovian Discriminator [9] to construct the main discriminator architecture. Our discriminator has two channels, where one is encode the video RGB space, and another is to encode the residual error.

the targeted ground-truth video sequence is $\{G_t|t = 1...n\}$. Let W_t denote the optical flow from frame I_{t-1} to I_t. In our experiment, we use the FlowNet2.0 [8] to compute the optical flow. We warp O_{t-1} to $\hat{O}_t$ via bilinear interpolation. Specifically, for each pixel $\boldsymbol{p} = (x, y)$ in the frame O_{t-1}, $\hat{O}_t(\boldsymbol{p}) = O_{t-1}(\boldsymbol{p} + W_t(\boldsymbol{p}))$. Now we can compute the forward residual error for the predicted frame O_t: $\hat{E}_t^o = O_t - \hat{O}_t$. For the ground-truth video sequence, we use the same method to get $\hat{E}_t^g = G_t - \hat{G}_t$. Noted the predicted video and ground-truth video use the same optical flow W_t.

To detect the occlusion between two frames, we perform a forward-backward consistency check of the optical flow. Let $W_t^f(\boldsymbol{p}) = (u, v)$ be the optical flow in forward direction from frame I_{t-1} to I_t and $W_{t-1}^b(\boldsymbol{p}) = (\hat{u}, \hat{v})$ the flow in backward direction from frame I_t to I_{t-1}. u and $\hat{u}$ are the displacement in x-coordinate, and v and $\hat{v}$ are the displacement in y-coordinate. Denote by $\tilde{W}$ the forward flow warped to the second image via bilinear interpolation:

$$\tilde{W}(\boldsymbol{p}) = W_t^f(\boldsymbol{p} + W_{t-1}^b(\boldsymbol{p})), \tag{1}$$

According to the study in the literature [19], we can mark as occlusions those areas where the following inequality holds:

$$|\tilde{W} + W_{t-1}^b|^2 > 0.01(|\tilde{W}|^2 + |W_{t-1}^b|^2) + 0.5, \tag{2}$$

According to the above inequality, we obtain D_t, which denotes the detected occlusion region for frame I_t.

Motion boundaries are detected using the following inequality:

$$|\nabla\hat{u}|^2 + |\nabla\hat{v}|^2 > 0.01|W_{t-1}^b|^2 + 0.002, \tag{3}$$

where $\nabla\hat{u}, \nabla\hat{v}$ are the gradient of the location $\boldsymbol{p}$ in the optical flow W_{t-1}^b. We let M_t denote the detected motion boundary for frame I_t. Finally, we obtain the modified residual error $E_t^o = D_t \times M_t \times \hat{E}_t^o$ for the predicted frame O_t, and $E_t^g = D_t \times M_t \times \hat{E}_t^g$ for the ground-truth frame G_t. In Fig.2, for the real data, the residual error set is $\mathcal{E}^g = \{..., E_t^g, E_{t+1}^g, ...\}$, and for the predicted data, the residual error set is $\mathcal{E}^o = \{..., E_t^o, E_{t+1}^o, ...\}$.

3.2 Network Architecture

3.2.1 RNN-based Generator. To consider the relation between adjacent frames, we combine the RNN with U-Net [17] to construct the Generator. The network architecture is as plot in Fig. 3. The U-Net contains an encode part and a decode part. The previous video frame is first input to the encode network to get the inner feature. This inner feature is propagated to the current frame and is combined with the inner feature of current frame. The combined feature is finally input the decode network to output the targeted frame.

3.2.2 Two-channel RNN-based Discriminator. For the discriminator, we also use a RNN-based network architecture. Inspired by the discriminator in [9], we combine RNN with the Markovian discriminator to construct our discriminator. Besides the traditional channel to encode the spatial relation in the video RGB space (called spatial channel), we use another channel to encode the residual error, which can be called temporal channel. These two channels have the same network architecture. The temporal channel encodes the temporal consistency for the entire video as a whole. Therefore, the two-channel discriminator ensures the global temporal consistency. The discriminator network is plotted in Fig. 4.

3.2.3 Details of The Network. Let Ck denote a Conv-BN-ReLU layer with k filters. CDk denotes a a Conv-BNDropout-ReLU layer with a dropout rate of 50%. All convolutions are 4×4 spatial filters applied with stride 2. RNN denotes the RNN layer, where it firstly concatenates the input and hidden units, and then applies two convolutions with $1\times$ 1 filter and stride 1 to compute the output unit and hidden unit, respectively.

The generator architecture is C64-C128-C256-C512-C512-C512-C512-C512-RNN-CD512-CD512-CD512-C512-C256-C128-C64. The left of RNN is the encoder network, and the right is the decoder network. All ReLUs in the encoder are leaky, with slope 0.2, while ReLUs in the decoder are not leaky.

The discriminator architecture is C64-C128-C256-C512-RNN-AvePool. The AvePool denotes that all the outputs of RNN for each frame are fed into a layer of average pooling to get a 1 dimensional output, followed by a Sigmoid function.

3.3 Global and Local Temporal Consistency

We combine the local and global temporal consistency to jointly ensure the video stability. For local consistency, as mentioned before, we compute the residual error between the predicted frames $\mathcal{E}^o =$

$\{..., E^o_t, E^o_{t+1}, ...\}$. In the training phase, the $\mathcal{E}^o = \{..., E^o_t, E^o_{t+1}, ...\}$ are the temporal consistency loss, we directly minimize them to achieve the local consistency. The temporal loss is computed as follows:

$$\mathcal{L}_{temporal} = \frac{1}{T}\sum_{t=1}^{T} ||E^o_t||_1 \tag{4}$$

where T is the number of frames in a video. For the global temporal consistency, we input the predicted output $\mathcal{E}^o = \{..., E^o_t, E^o_{t+1}, ...\}$ and the ground-truth results $\mathcal{E}^g = \{..., E^g_t, E^g_{t+1}, ...\}$ to the two-channel discriminator, and then compute the loss of GAN as follows:

$$\begin{aligned}\mathcal{L}_{cGAN} &= \mathbb{E}_{I,O\sim p_{data}(I,O)}[logD(I,O)] \\ &+\mathbb{E}_{I\sim p_{data}(I),z\sim p_z(z)}[log(1-D(I,G(I,z)))],\end{aligned} \tag{5}$$

where I is the input videos, O is the predicted videos, and z is the random noise vector. Eq.5 is the loss of a conditional GAN. As plotted in Fig. 2, the two-channel discriminator encodes the global temporal consistency. Therefore, the global temporal consistency is achieved by minimizing the Eq.5.

Like [9], we also add the L1 distance between the predicted frames and the ground-truth frames as follows:

$$\mathcal{L}_{spatial} = \frac{1}{T}\sum_{t=1}^{T} ||O_t - G_t||_1 \tag{6}$$

Overall, our final objective is as follows:

$$\mathcal{L} = \mathcal{L}_{temporal} + \lambda\mathcal{L}_{spatial} + \eta\mathcal{L}_{cGAN} \tag{7}$$

where λ and η are the parameters. To optimize the problem, we follow the standard approach from: we alternate between one gradient descent step on D, then one step on G. We apply the Adam solver [13].

4 EXPERIMENTS

To verify the proposed video-to-video translation framework, we test it on three different computer vision tasks: video segmentation, video colorization, and video super-resolution.

4.1 Datasets

Video Segmentation: We conduct experiments on the DAVIS 2017 dataset [15]. DAVIS 2017 is proposed for the video segmentation, which consists of 90 high-quality videos. All the frames come with high-quality per-pixel annotation of the foreground object, from which 60 are taken for training and 30 for validation. We use the subsampled version with a resolution of 854×480 pixels. In the training phase, we sample each 5 frames to obtain the video clips. In this way, we obtain the training set with 903 video clips. In the testing phase, the original videos are used to accomplish the video-to-video translation.

Video Colourization and Video Super-resolution: We also use the DAVIS 2017 to construct the training set and testing set for video colourization and video super-resolution. For video colourization, we convert the original videos to their gray versions, and thus obtain the color and gray video pairs. For video super-resolution, we first reduce the resolution for each original video frame with 1/2 reduction, and then compute their ×2 super-resolution videos with bicubic interpolation. In this way, we get the low-resolution and super-resolution video pairs.

4.2 Evaluation metrics

Video Segmentation: We use the evaluation metrics introduced in [15], i.e., Region Similarity $\mathcal{J}$, Contour Accuracy $\mathcal{F}$, and Temporal stability $\mathcal{T}$. Specifically, $\mathcal{J}$ is defined as the intersection over-union of the estimated segmentation M and the ground-truth mask G, $\mathcal{J} = \frac{M\cap G}{M\cup G}$. For contour accuracy, $\mathcal{F} = \frac{2P_cR_c}{P_c+R_c}$, where P_c and R_c denote the contour-based precision and recall between the contour points of $c(M)$ and $c(G)$, respectively. Temporal stability $\mathcal{T}$ is to quantitatively measure the temporal consistency between two frames. $\mathcal{T}$ is firstly proposed in [15], which is computed by the so-called mean cost per matched point. For details, please see the cited paper. For $\mathcal{T}$, a lower value is better.

Video Colourization and Video Super-resolution: To measure the similarity between predicted videos and the ground-truth videos, we use the PSNR (Peak Signal to Noise Ratio). PSNR is an approximation to human perception of reconstruction quality. A higher PSNR generally indicates that the predicted video frame is of higher quality.

4.3 Results and Discussions

We compare with four different experimental settings and report their corresponding results. The first one is the image-to-image method [9], where we regard all the video frames in the dataset as discrete images and train the model. In the testing phase, the trained image2image model is performed on each video frame independently, and then all the predicted frames are concentrated into a video. The second one is the proposed video-to-video framework introduced in section 3.2, while giving up the global and local temporal consistency. The third one is adding the local temporal consistency on the basis of the second one. The forth one is the whole proposed framework, i.e., the video-to-video method with simultaneous global and local temporal consistency.

The final experimental results are listed in Table 1. From the table, we can draw the following conclusions: 1. The image2image method shows the worst performance on all the three tasks. This is reasonable because image2image is designed for image translation. The temporal interactions between frames are not considered, and thus it is not suitable for the video-to-video translation. 2. As a contrast, the video2video uses the RNN to encode the temporal interactions between frames, and obtain a remarkable improvement in the evaluation metrics over the image2image. 3. To ensure the temporal consistency, we add the temporal loss (local temporal consistency) into the optimization process. The table shows that video2video+local obtains a slight improvement over the video2video method, which demonstrates the limited power of temporal loss in the supervised tasks. Noted, in the super-resolution task, the performance of video2video+local is even worse than the video2video. This shows that although local temporal consistency works well in the unsupervised task like video stylization, it is not suitable to the supervised tasks like video segmentation, video super-resolution. The reason for this is that in the supervised tasks, minimizing the loss between predicted frames and the ground-truth frames has made the video stable, which limits the effect of temporal loss. While in the unsupervised tasks, the ground-truth data is absent. In this case, the temporal loss is important to the temporal consistency. 4. The proposed global+local consistency obtains the

Table 1: The performance of the proposed video-to-video framework in different settings and tasks on DAVIS2017 dataset.

Tasks	Segmentation			Gray2color	Super-Resolution
Models	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{T}$	$\mathcal{P}$	$\mathcal{P}$
image2image	0.5066	0.4617	0.3494	33.0289(dB)	30.1813(dB)
video2video	0.7290	0.7813	0.2006	33.3417(dB)	30.7219(dB)
video2video+local	0.7359	0.7932	0.1767	34.8182(dB)	30.5414(dB)
video2video+local+global	**0.9229**	**0.9589**	**0.0537**	**37.8843**(dB)	**33.0043**(dB)

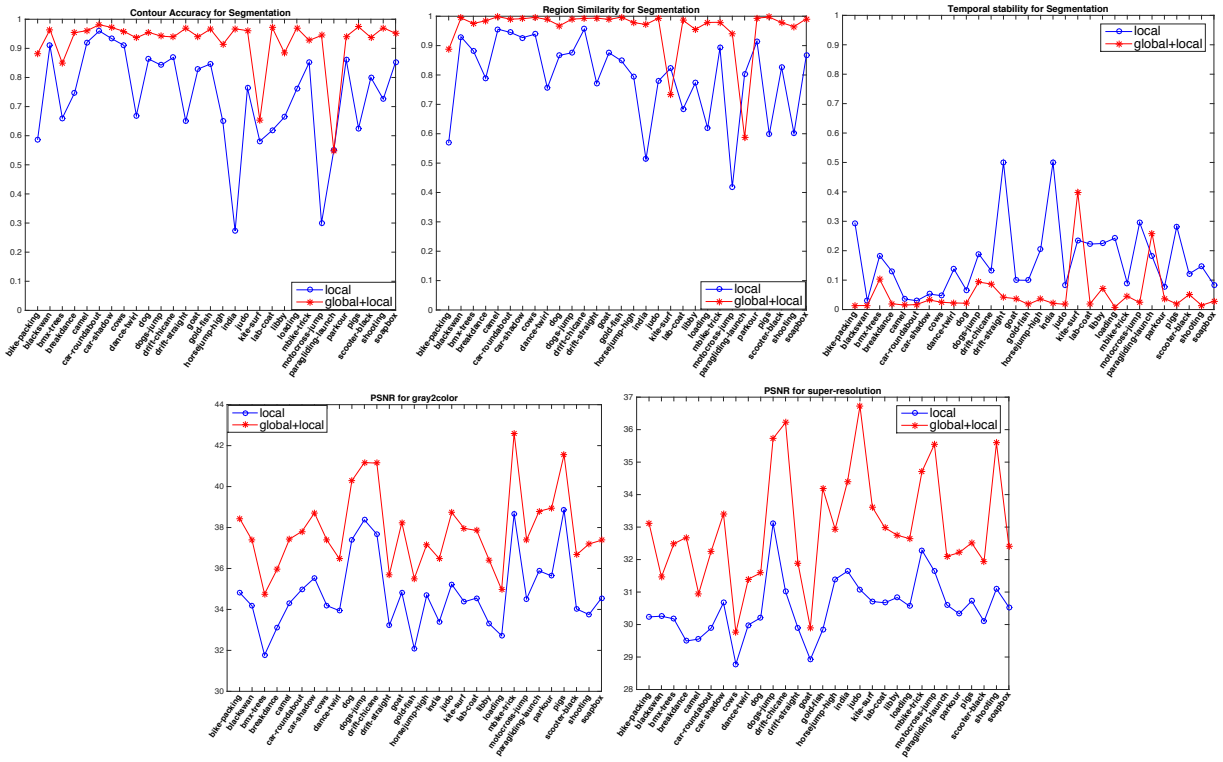

Figure 5: The performance for each class in the testing set versus local and global+local consistency, respectively. The used dataset has 30 classes in the testing set, we list these 30 classes in the x-coordinate, and their corresponding performance in the y-coordinate. Noted for the temporal stability, a lower value is better.

best performance, and furthermore, achieves a major improvement (Specifically, 0.187 versus $\mathcal{J}$, 0.1657 versus $\mathcal{F}$, and 0.0239 versus $\mathcal{T}$ for segmentation. For gray2color and super-resolution, it adds 3.06 dB and 2.46 dB, respectively), which shows the effectiveness of our method.

We also give the performance for each class in the testing set in Fig. 5. For each figure, the performance of local and global+local are illustrated. Fig. 5 also shows advantage of the proposed global+local consistency.

In addition, the qualitative results on the three tasks are given in Fig. 6-8. In each figure, besides the original frames, we list the results of image2image, video2video+local consistency, and video2video+global consistency, respectively. Form each figure, we see that on all the three tasks, the temporal inconsistency of the predicted frames of image2image method are obvious and easily captured by human eyes. With the adding local and global temporal consistency, the video quality is increasing remarkably. Specifically, In Fig. 6, for image2image method, the tails of goose are not segmented completely and each frame shows slightly different appearance. This is reasonable because the segmentation is performed by frames. As a comparison, the outputs with local consistency are more accurate owing to the frame-wise temporal consistency. However, there still exits some mutation like the head of goose in the third frame. With the adding global temporal consistency, the segmentation results are further refined, and the appearance of varying frames are smooth. Figure. 7-8 also show the same trend (see the color in Fig. 7, and the image quality in Fig. 8). Our video2video framework achieves a major improvement on all the three tasks, which demonstrates that the proposed architecture can learn a good mapping between the input video and the output video, and is regardless of the types of the video content.

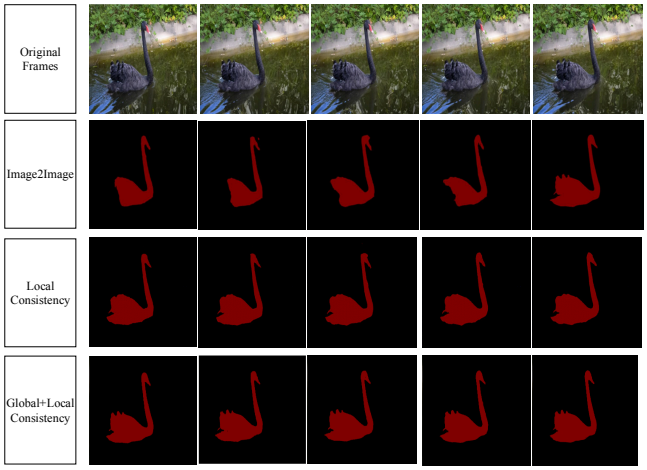

Figure 6: The qualitative results on the video segmentation versus image2image, video2vide+local consistency, and video2video+global consistency. The original frames are listed in the top row.

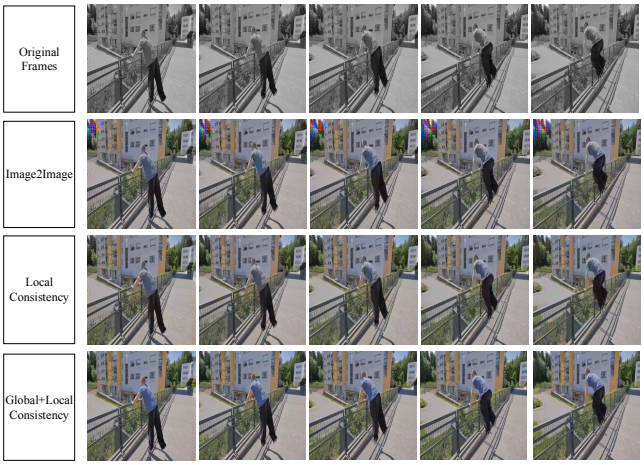

Figure 7: The qualitative results on the video colourization versus image2image, video2vide+local consistency, and video2video+global consistency. The original frames are listed in the top row.

5 CONCLUSIONS

In this paper, we proposed an unified video-to-video translation framework to accomplish three tasks: video super-resolution, video colourization, and video segmentation. To overcome the temporal inconsistency and obtain stable videos, a residual error based mechanism was presented to ensure the local and global consistency of the same location in different frames. Our method is based on the widely used GAN framework. We adapted the video generation into this framework and integrated the residual error based mechanism. Extensive experiments conducted on different video-to-video translation tasks verified the effectiveness and flexibleness of the proposed method.

Figure 8: The qualitative results on the video super-resolution versus image2image, video2vide+local consistency, and video2video+global consistency. The original frames are listed in the top row.

ACKNOWLEDGMENTS

This work was supported by NSFC Projects (Nos. 61571261, 61620106010, 61621136008, 61332007, and U1611461), Beijing NSF Project (No. L172037), MIIT Grant of Int. Man. Comp. Stan (No. 2016ZXFB00001), the Youth Top-notch Talent Support Program, Tiangong Institute for Intelligent Computing, NVIDIA NVAIL Program, Siemens, and partially funded by Microsoft Research Asia and Tsinghua-Intel Joint Research Institute.

REFERENCES

[1] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2015. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 196.

[2] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent Online Video Style Transfer. *IEEE International Conference on Computer Vision (ICCV)* (2017).

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2018), 834–848.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*. 184–199.

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2414–2423.

[7] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. Real-time neural style transfer for videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*. IEEE, 7044–7052.

[8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[10] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. 2017. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. *IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR) (2017).

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

[12] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. 2013. State of the" Art?: A Taxonomy of Artistic Stylization Techniques for Images and Video. *IEEE Transactions on Visualization and Computer Graphics* 19, 5 (2013), 866–885.

[15] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675* (2017).

[16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. *ICML* (2016).

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[18] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *German Conference on Pattern Recognition*. Springer, 26–36.

[19] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. 2010. Dense point trajectories by GPU-accelerated large displacement optical flow. In *European conference on computer vision (ECCV)*. Springer, 438–451.

[20] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. 2017. Detail-revealing Deep Video Super-resolution. *IEEE International Conference on Computer Vision (ICCV)* (2017).

[21] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. 2010. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* 19, 11 (2010), 2861–2873.

[22] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint* (2017).

[23] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*. Springer, 649–666.

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision (ICCV)* (2017).