Unsupervised Coherent Video Cartoonization with Perceptual Motion Consistency

Zhenhuan Liu[1,2], **Liang Li**[1*], **Huajie Jiang**[3]
Xin Jin[3], **Dandan Tu**[3], **Shuhui Wang**[1], **Zheng-Jun Zha**[4]
[1] Institute of Computing Technology, [2] University of Chinese Academy of Sciences
[3] Huawei Technologies, [4] University of Science and Technology of China
zhenhuan.liu@vipl.ict.ac.cn, liang.li@ict.ac.cn, {jianghuajie1, jinxin11, tudandan}@huawei.com,
wangshuhui@ict.ac.cn, zhazj@ustc.edu.cn

Abstract

In recent years, creative content generations like style transfer and neural photo editing have attracted more and more attention. Among these, cartoonization of real-world scenes has promising applications in entertainment and industry. Different from image translations focusing on improving the style effect of generated images, video cartoonization has additional requirements on the temporal consistency. In this paper, we propose a spatially-adaptive semantic alignment framework with perceptual motion consistency for coherent video cartoonization in an unsupervised manner. The semantic alignment module is designed to restore deformation of semantic structure caused by spatial information lost in the encoder-decoder architecture. Furthermore, we devise the spatio-temporal correlative map as a style-independent, global-aware regularization on the perceptual motion consistency. Deriving from similarity measurement of high-level features in photo and cartoon frames, it captures global semantic information beyond raw pixel-value in optical flow. Besides, the similarity measurement disentangles temporal relationships from domain-specific style properties, which helps regularize the temporal consistency without hurting style effects of cartoon images. Qualitative and quantitative experiments demonstrate our method is able to generate highly stylistic and temporal consistent cartoon videos.

Introduction

Cartoon movies are very popular and attractive across the world, but creating even a short cartoon movie involves a complex and time-consuming production process with multiple stages. Recently, deep learning based methods have developed a lot of techniques for creative content generation, like style transfer, semantic image synthesis and neural photo editing. Video cartoonization aims at creating coherent cartoon-styled videos based on real-world videos, as shown in Figure 1. As a practical technique, it has promising applications in entertainment and industry.

Given two collections containing unpaired photo and cartoon images, cartoonization usually leverages generative adversarial networks(GANs) to map the input photo into the distribution of cartoon images. Different from traditional style transfer which often adds textures like brush strokes, cartoonization aims to create highly conceptual and abstract images. Besides, due to the dynamic training characteristic of unsupervised GANs, it's harder to achieve stable convergence of style effects in generated images than style transfer which has explicit style constraints. Compared to image cartoonization, video cartoonization has higher requirements on the temporal consistency between input and output frames. There are two challenges to the coherence of output video. The first one is keeping the structure consistency between each input and output frame. It's difficult to preserve structure consistency of output image while depicting the highly abstracted cartoon image structure in unsupervised learning. The second challenge is to guarantee the temporal coherence of generated frames. Previous works have shown that the output of deep neural networks is very sensitive to small changes in pixel values (Azulay and Weiss 2019). The generator of video cartoonization needs to be robust under different transformations of input frames.

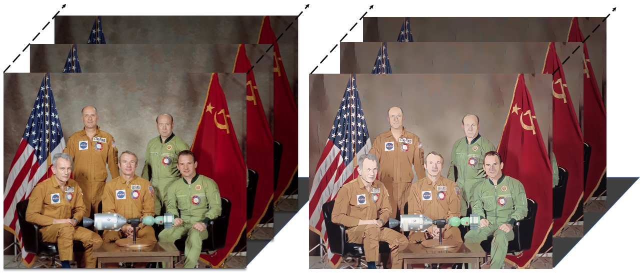

Figure 1: Illustration of video cartoonization. It requires smooth surface and clear border with temporal consistency for a typical cartoon artwork.

There are mainly three kinds of methods to solve the problem of video temporal inconsistency. The first method explicitly employs optical flow estimated between input frames to warp the last output frame for generation (Chen et al. 2017). However, this method highly relies on the accuracy of flow estimation, which is difficult for complex real-world environments. The second one is to build a task-independent model that can repair temporal inconsistency of videos generated by different image-translation models, such as Deep Video Prior(Lei, Xing, and Chen 2020). But

[*] Corresponding author.

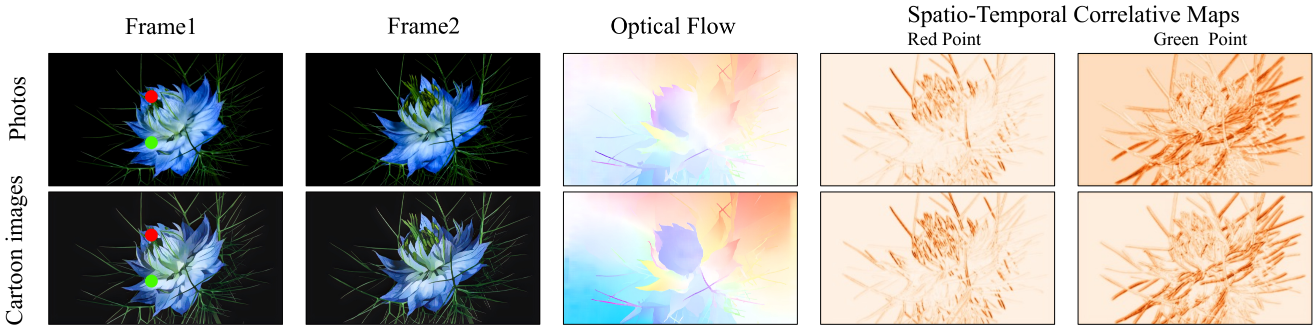

Figure 2: Optical flow and our spatio-temporal correlative maps for two real world photos and its cartoon images. We select the red and green points as source points and build their spatio-temporal correlative maps. The spatio-temporal correlative maps are more robust under different domains.

this method is time-consuming and hard for real-time applications. Another method is to add a temporal regularization upon image-to-image translation models at training stage, which encourages the coherence of output frames (Wang et al. 2020b). These traditional methods for temporal consistency often suffer from the limitations of pixel-wise motion representation, e.g., optical flow. Since images in different domains have different properties in appearances and textures, optical flow estimated in the photo domain may not align well with the cartoon domain. As shown in the third column of Figure 2, the optical flows estimated for real images and cartoon images show obvious discrepancies, especially for the central region with large distortions. Further, pixel-wise motion representation is unable to handle disocclusion problems where newly appeared pixels have no correspondence with last frame.

To solve above problems, we propose a spatially-adaptive semantic alignment framework with perceptual motion consistency for coherent video cartoonization in an unsupervised manner. The Spatially-adaptive Semantic Alignment(SSA) module is designed to restore the local shift and deformation of semantic structure caused by spatial information lost in the traditional encoder-decoder network architecture. To overcome the limitations of dense motion representations for complex scenes with large displacements and disocclusion under different domains, we propose a style-independent global-aware regularization on Perceptual Motion Consistency(PMC) via spatio-temporal correlative maps between input and output video.

Specifically for the SSA module, we first calculate the semantic distance of feature vectors between encoder and decoder at each scale. After that, we relocate the decoder's feature in a local patch to the position of its nearest encoder's features, resulting in a refined structure corresponding to the encoder's structure. Such that the structure consistency between input and output image can be better preserved. For the PMC module, we introduce spatio-temporal correlative maps to impose regularization on perceptual motion consistency. Different from the pixel-wise hard correspondence of optical flow, it measures feature similarity instead of absolute pixel value for each patch of current frame with next frame. This map builds a probabilistic distribution of each object's flow direction in a global context, which can be generalized in complex situations with disocclusion. Furthermore, the similarity measurement disentangles the temporal relationship from domain-specific attributes such as color, lighting and texture, so that we can formulate a style-independent motion representation for different domains.

Our contributions can be summarized as follows: (1) We propose an effective framework with a style-independent global-aware regularization on perceptual motion consistency to generate temporal consistent cartoon videos. (2) A novel generator with spatially-adaptive semantic alignment is designed to generate semantic-aware structure consistent cartoon images. (3) We conduct detailed qualitative and quantitative experiments and demonstrate our method achieves both stylistic cartoon effect and temporal consistency. Our code will be released on github.

Related Works

Unsupervised Image-to-Image Translation aims to learn the mapping from a source image domain to a target image domain with unpaired data. One of the key challenges is to build a meaningful structure correspondence between input and output images. Cycle-consistency and its variants are often used to guarantee the output image can reconstruct the input image (Zhu et al. 2017; Huang et al. 2018; Liu, Breuel, and Kautz 2017; Lin et al. 2020a,b). (Park et al. 2020) incorporated patchwise contrastive learning to achieve structure preservation. Recently (Zheng, Cham, and Cai 2021) proposed spatially-correlative maps as domain-invariant representations to restrict structure consistency. In this paper, we learn the unsupervised translation from photo domain to cartoon domain. Different from the above methods only applying to images, we extend to generate videos with temporal consistency.

Image Cartoonization aims to generate cartoon images with clear edges, smooth color shading and relatively simple textures from real-world photos. Originally, the pioneering work CartoonGAN (Chen, Lai, and Liu 2018) was proposed for image cartoonization, which introduced a novel adversarial loss to encourage clear edges. Recently, (Wang and Yu 2020) developed three white-box image representations that reflect different aspects of cartoon images, including texture, surface and structure. Although these models can generate cartoon styled videos by applying per-frame translation, their output videos show temporal inconsistencies and flickering artifacts.

Video Temporal Consistency is a research topic about solving the flickering problem when applying different kinds of image-based models to videos. Task-independent methods design a single model for different tasks, which generate a coherent video from separately processed frames. (Lai et al. 2018) utilized FlowNet2 to estimate optical flow as temporal loss to train the transformation network. (Lei, Xing, and Chen 2020) leveraged deep video prior to iteratively optimize each sequence. Task-specific approaches develop different strategies according to each domain, such as designing specific network architectures (Tassano, Delon, and Veit 2020; Deng et al. 2021) or embedding optical flow estimation to capture information of motion (Chen et al. 2017). Recently, (Wang et al. 2020b) proposed compound regularization for temporal consistent video style transfer. These methods often heavily rely on pixel-wise correspondence which is sensitive to the different appearances of objects in photo and cartoon domain.

Methodology

Overview

Given a real world video composed of frames $\{s^0, s^1, ...s^n\}$, video cartoonization aims to generate corresponding cartoon frames $\{t^0, t^1, ...t^n\}$ whose temporal consistency is preserved. As shown in Figure 3, at training stage, our model transforms input consecutive real world photos s^0 and s^1 into corresponding cartoon images t^0 and t^1. The generator builds upon an encoder-decoder architecture consisting of downsampling, residual blocks and upsampling layers. We introduce a spatially-adaptive semantic alignment module into the generator to render semantic-aware structure consistent cartoon images. To restrict perceptual motion consistency, we first extract multi-level features from both input and cartoon images with pretrained deep network. Then, we introduce the spatio-temporal correlative maps of features to regularize perceptual motion consistency between input and output video. We employ adversarial mechanism to optimize the cartoon effect of generated images, where the generator and discriminator are updated alternatively.

Spatially-Adaptive Semantic Alignment

To address the problem of structure deformation in the decoding stage, we introduce the spatially-adaptive semantic alignment(SSA) module into the generator. As shown in Figure 3, to refine the structure of upsampled feature maps $g^l(x) \in \mathbb{R}^{C\times H\times W}$ in the l-th level of decoder, we adopt the feature maps $f^l(x)$ of encoder in corresponding layer to relocate the placement of decoder features.

Specifically, for a source feature vector $f_i^l(x)$ in location i, we define a local patch of size $N = R^2$ around i in $g^l(x)$ as its semantic candidates $\{g_j^l(x)|j \in \{1, 2..., N\}\}$. We first compute the magnitude of similarity between them:

$$z_{ij}^l(x) = (f_i^l(x))^T(g_j^l(x)) \tag{1}$$

Then, we apply spatial-wise softmax to $[z_{i1}^l, z_{i2}^l, ...z_{iN}^l]$, which constructs a normalized correlative map as a kernel. Each element of the kernel is calculated as follows:

$$\alpha_{ij}^l(x) = \frac{exp(z_{ij}^l(x))}{\sum_{k=1}^N exp(z_{ik}^l(x))} \tag{2}$$

Finally, we obtain the refined feature in location i by aggregating semantic candidates with above kernel weight:

$$r_i^l(x) = \sum_{i=1}^N \alpha_{ij}^l(x) g_j^l(x) \tag{3}$$

Similar to pooling operations, this module replaces the output of convolution with a summary statistic of the nearby outputs. Furthermore, there are two properties of our proposed module: First, it's spatially-adaptive, where different kernels are applied for different locations. Second, instead of embedding hand-craft static pooling like max or average pooling, our module dynamically generates the kernel based on the semantic similarity of features.

For pixel-wise prediction tasks like image translation and semantic segmentation, the structure of output image needs to align with input image. Suffering from the low-resolution bottleneck, encoder-decoder architectures lacks fine-grained structure alignment. Traditional methods like U-net often use skip connections to alleviate this problem. The skip connection propagates encoder's high-resolution information to upsampled output with concatenation for precise localization. However, it only strengthens the spatial alignment instead of semantic alignment. By contrast, our proposed SSA repairs the semantic misalignment by relocating each local patch of upsampled outputs with the high-resolution feature maps of encoder(The influence of the patch width will be further discussed in ablation study.).

Besides, the SSA has the edge-preserving smoothing characteristic especially suitable for cartoonization task. For features within the same semantic region, the kernel of semantic candidates tends to be an average kernel, which helps render smooth surface in this region. For features near the border of a semantic region, the kernel will assign dominant weight to locations within the border. Thus the edges between different objects can be well preserved, which is consistent with cartoon image's property, as shown in Figure 1.

Perceptual Motion Consistency

Traditional methods often utilized warping error with optical flow to inhabit temporal inconsistency between output frames (Chen et al. 2017; Park et al. 2019; Kim et al. 2019). The formulation of warping error is:

$$\mathcal{L}_{\text{warp}} = \|t^i - W_{i-1,i}(t^{i-1})\| \tag{4}$$

where $W_{i-1,i}$ denotes the optical flow that warps pixels of s^{i-1} to s^i.

However, there are three limitations of the pixel-wise temporal regularization. First, optical flow describes the spatial variation of pixel but ignores its value variation. Actually, the brightness and color appearances of the same object might be different on consecutive frames due to illumination

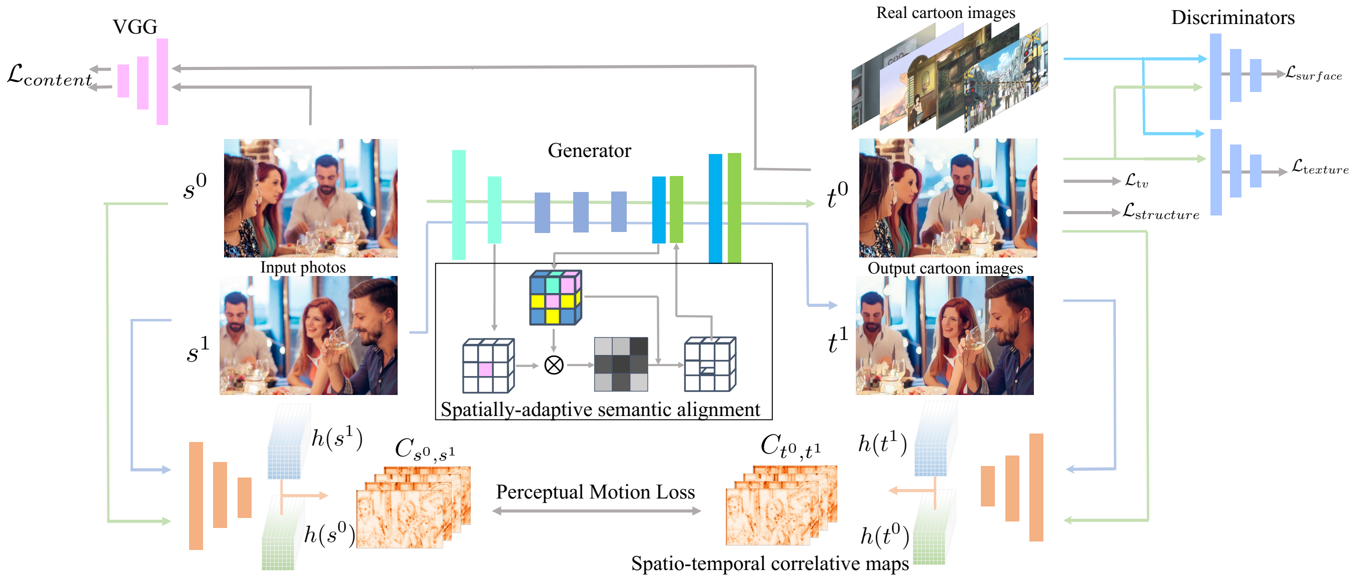

Figure 3: The architecture of our proposed method. Given two consecutive frames of real-world photos, the spatially-adaptive semantic alignment module is introduced into our generator to produce semantic-aware structure consistent cartoon images. After that we derive the spatio-temporal correlative maps from extracted features of input and cartoon images to regularize the perceptual motion consistency. Besides, content loss is used to restrict the structure correspondence between input and output image, and adversarial loss is used to optimize the cartoon effect of output images.

effects. As a consequence, warping error tends to keep the same value corresponding to last frame's output, even when the input value has changed. Second, images in different domains have different properties in appearances and textures, optical flow estimated in the photo domain is hard to align with that in cartoon domain, which is illustrated in Figure 2. Third, optical flow can not handle the disocclusion problem since newly appeared pixels have no correspondence with last frame.

Here, we propose the perceptual motion consistency to formulate a style-independent global-aware motion representation under photo and cartoon domains. We denote h as the pretrained feature extractor. Take input video as example, for the extracted feature $h(s^i)_j$ of patch j in current frame s^i, we compute its similarity magnitudes with all features of next frame. We introduce it as the *spatio-temporal correlative map*, formally:

$$C_{s^i_j,s^{i+1}} = \left(h\left(s^i\right)_j\right)^T \left(h\left(s^{i+1}\right)\right) \tag{5}$$

where $h\left(s^i\right)_j \in \mathbb{R}^{M\times 1}$, M is the number of channels in feature maps, $h\left(s^{i+1}\right) \in \mathbb{R}^{M\times N}$ contains features for all locations of size N in frame s^{i+1}, such that $C_{s^i_j,s^{i+1}} \in \mathbb{R}^{1\times N}$ captures the correlation between source patch j with each patch of next frame.

Next, we formulate the perceptual motion representation between consecutive frames as a collection of spatio-temporal correlative maps:

$$C_{s^i,s^{i+1}} = [C_{s^i_0,s^{i+1}}; C_{s^i_1,s^{i+1}}; ...; C_{s^i_N,s^{i+1}}] \in \mathbb{R}^{N\times N} \tag{6}$$

After that, we calculate the multi-level spatio-temporal correlative maps corresponding to extracted features at different scales between source and target domain. Here cosine loss is used to restrict the perceptual motion consistency:

$$\mathcal{L}_{\text{motion}} = \|1 - \cos(C_{s^i,s^{i+1}}, C_{t^i,t^{i+1}})\|_1 \tag{7}$$

The spatio-temporal correlative maps model the motion as correspondence of semantic features in a global context. It's more robust under domains with different appearances and styles by exploiting the similarity measurement of high-level features. Besides, it can handle the problem of disocclusion by leveraging the global relationships of newly appeared objects.

Loss Functions

Besides the perceptual motion loss, the cartoonization effect of generated images is also optimized with adversarial loss. Here, we model three representations extracted from images to optimize the generated cartoon images, including the surface representation that contains smooth surface of cartoon images, the structure representation that refers to flattened global content, the texture representation that reflects textures and details in cartoon images.

To derive surface representation, guided filter $\mathcal{F}_{gf}$ (He, Sun, and Tang 2013) is adopted for edge-preserving filtering. It removes the textures and details of input image with the help of a guided image, which can be the input image itself. A discriminator D_s is leveraged to determine whether the structure representations of output images are similar to that of cartoon images. Let I_p denote the input photo and I_c

indicate the reference cartoon images, the surface loss is

$$\begin{aligned}\mathcal{L}_{\text{surface}}(G, D_s) &= \log D_s(\mathcal{F}_{gf}(\boldsymbol{I}_c, \boldsymbol{I}_c)) \\ &+ \log(1 - D_s(\mathcal{F}_{gf}(G(\boldsymbol{I}_p), G(\boldsymbol{I}_p))))\end{aligned} \tag{8}$$

To derive structure representation, a superpixel algorithm (Felzenszwalb and Huttenlocher 2004) is deployed to segment images into separate semantic consistent regions. After that, each region is filled with a corresponding color, which generates the structure representation $\mathcal{F}_{st}$, The structure loss is:

$$\mathcal{L}_{\text{structure}} = \|VGG(G(\boldsymbol{I}_p)) - VGG(\mathcal{F}_{st}(G(\boldsymbol{I}_p)))\| \tag{9}$$

To derive texture representation, we convert images into grayscale. Another discriminator D_t is used to distinguish the distribution of cartoon images from generated images. And the texture loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{texture}}(G, D_t) &= \log D_t(\mathcal{F}_{gray}(\boldsymbol{I}_c)) \\ &+ \log(1 - D_t(\mathcal{F}_{gray}(G(\boldsymbol{I}_p))))\end{aligned} \tag{10}$$

Besides, content loss is utilized to regularize the structure consistency between photo and cartoon images,

$$\mathcal{L}_{\text{content}} = \|VGG\left(G\left(\boldsymbol{I}_p\right)\right) - VGG\left(\boldsymbol{I}_p\right)\| \tag{11}$$

The total variation loss is incorporated to encourage the smoothness of generated images,

$$\mathcal{L}_{tv} = \frac{1}{H * W * C}\|\nabla_x\left(G\left(\boldsymbol{I}_p\right)\right) + \nabla_y\left(G\left(\boldsymbol{I}_p\right)\right)\| \tag{12}$$

Finally, the objective of full model is formulated as the summation of adversarial loss, perceptual motion loss, content loss and total variation loss:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \lambda_1\mathcal{L}_{\text{surface}} + \lambda_2\mathcal{L}_{\text{texture}} + \lambda_3\mathcal{L}_{\text{structure}} \\ &+ \lambda_4\mathcal{L}_{\text{content}} + \lambda_5\mathcal{L}_{tv} + \lambda_6\mathcal{L}_{\text{motion}}\end{aligned} \tag{13}$$

Experiment

Experimental Setup

Dataset. For real-world photos, we adopt 10000 human face images from FFHQ dataset and 6227 landscape images from CycleGAN dataset (Zhu et al. 2017). For cartoon images, we use images from WhiteboxGAN (Wang and Yu 2020), including 10000 images of cartoon faces from P.A.Works, Kyoto animation and 14615 images from cartoon movies produced by Shinkai Makoto, Hosoda Mamoru, and Miyazaki Hayao. We apply random affine transformations on photo images to imitate input consecutive video frames. For test set, we use 1541 real-world images from (Zhu et al. 2017; Chen, Liu, and Chen 2020) and 1583 cartoon images from above cartoon movies. During training, all images are resized to 256 × 256 resolution.

Implementation Details. We implement our model using Pytorch Lightning (Falcon 2019). Our generator consists of two downsampling layers, four residual blocks and two upsampling layers. LeakyReLU is used as activation function. The local patch width of SSA is set as $R = 3$. Patch discriminator with spectral normalization (Miyato et al. 2018) is adopted to identify each patch's distribution. We use Adam (Kingma and Ba 2015) optimizer with momentums 0.5 and 0.99. The learning rate is set to 0.0002. The loss weight are set as $\lambda_1 = 0.1, \lambda_2 = 1, \lambda_3 = 200, \lambda_4 = 200, \lambda_5 = 20000, \lambda_6 = 0.1$.

Methods	FID to Cartoon	FID to Photo
CycleGAN	99.39	99.32
CUT	93.07	105.64
LSeSim	95.26	87.66
CartoonGAN Shinkai	98.46	70.86
CartoonGAN Paprika	111.40	136.12
CartoonGAN Hosoda	98.07	136.12
CartoonGAN Hayao	107.25	119.81
WhiteboxGAN	90.98	50.69
Ours	**87.96**	**41.11**

Table 1: Performance comparison of different methods. Smaller FID values indicates closer distance between generated images and reference images.

Evaluation of Cartoon Effects

In quantitative experiments, following recent works (Wang and Yu 2020; Park et al. 2020), we employ the Frechet Inception Distance (FID) (Heusel et al. 2017) to evaluate the quality of generated images. The FID to cartoon images reflects the quality of cartoon style effects, and the FID to photo indicates content consistency between input photo and generated images. We compare with five SOTA methods, including unsupervised image translation methods CycleGAN (Zhu et al. 2017), CUT (Park et al. 2020), LSeSim (Zheng, Cham, and Cai 2021) and image cartoonization methods CartoonGAN (Chen, Lai, and Liu 2018) and WhiteboxGAN (Wang and Yu 2020).

The quantitative results are shown in Table 1. We can observe that our model has the lowest FID to cartoon images. Besides, our model surpasses previous methods in FID to photo images by a large margin(9.58). This benefits from that our model can better preserve the structure information in source image, which proves the effectiveness of our SSA module.

The qualitative results are shown in Figure 4. We find that the generated images of our method depict vivid cartoon styles. First, from the aspects of color style, our result displays brighter lightness and higher image contrast without damaging the overall style of source image. Other SOTA methods like CUT, LSeSim and CartoonGAN generate images with severe color variation. Second, our method removes negligible details and presents a smoother surface for each semantic region. By contrast, compared methods especially for CycleGAN, CUT and CartoonGAN still keep original high-frequency information. Third, with the help of SSA module, our method highly abstracts source image and preserves semantic-aware consistent structure. Compared with the previous best model WhiteboxGAN, our method achieves more natural structure abstraction as shown by the man face in the fourth row.

Evaluation of Temporal Consistency

To evaluate the coherence of generated videos, we calculate the widely-used warping error as (Lei, Xing, and Chen 2020). For each output frame t^i of a sequence, we calculate

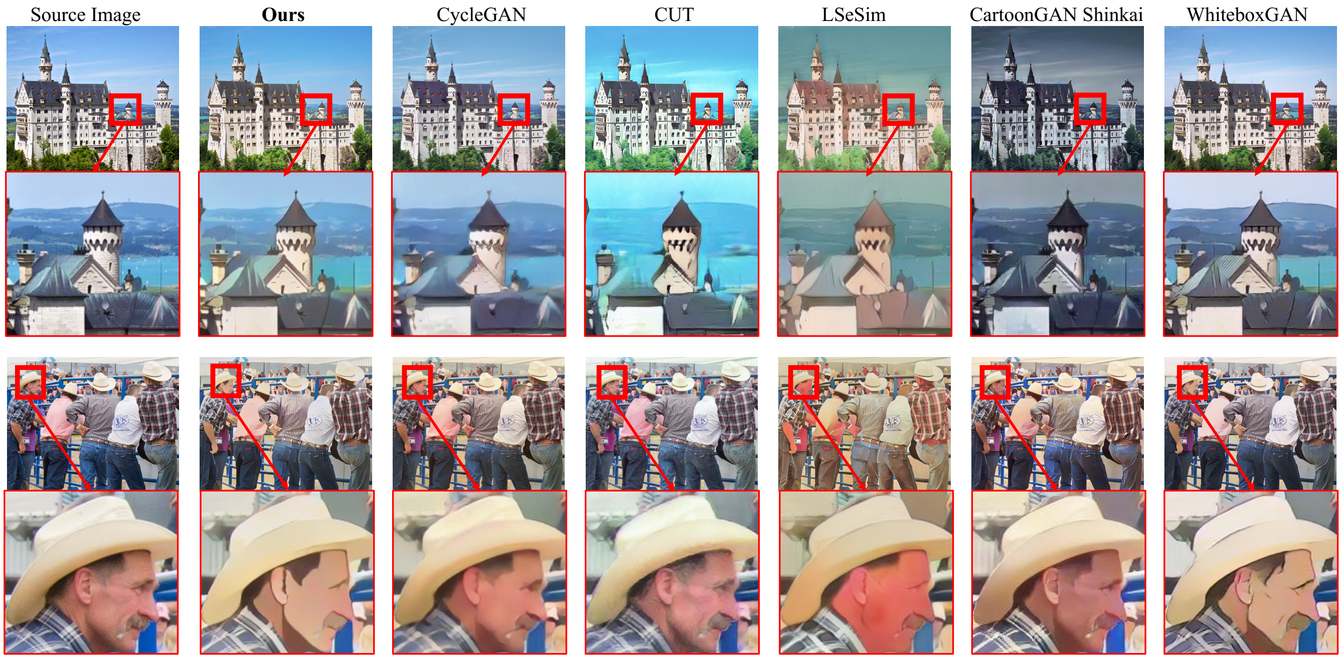

Figure 4: Qualitative comparison on cartoon effects, the first and third rows show the source image and generated cartoon images of different methods. The second and fourth rows refer to the magnified details in the red box of images.

its warping error with frame t^{i-1} for short-term consistency:

$$E_{short} = \frac{1}{N-1}\sum_{i=2}^{N}\|M \circ (W_{s^i \rightarrow s^{i-1}}(t^i) - t^{i-1})\|_1 \quad (14)$$

where $W_{s^i \rightarrow s^{i-1}}$ is the backward optical flow between s^{i-1} and s^i, M is the corresponding occlusion mask. We also calculate the warping error of each frame with frame t_1 for long-term consistency:

$$E_{long} = \frac{1}{N-1}\sum_{i=2}^{N}\|M \circ (W_{s^i \rightarrow s^1}(t^i) - t^1)\|_1 \quad (15)$$

Following recent works, we evaluate the temporal consistency on DAVIS (Perazzi et al. 2016) dataset. The optical flow is estimated with pretrained RAFT model (Teed and Deng 2020). In addition to aforementioned methods, we also adopt post-processing models including Blind (Lai et al. 2018) and Deep Video Prior(DVP) (Lei, Xing, and Chen 2020) upon the SOTA image cartoonization method WhiteboxGAN for comparison.

The quantitative results are shown in Table 2. Our method performs best for long-term temporal consistency and achieves lower short-term warping error than that of source video. For image translation models, CartoonGAN, LSeSim, CUT and WhiteboxGAN have higher warping error. For post-processing methods, Blind can decrease the short-term warping error but increase long-term warping error. DVP achieves the lowest warping error but it requires several minutes to process a single sequence. Our method has comparable performance on temporal consistency with post-processing methods, and this proves the capability of PMC to restrict temporal consistency.

Method	$E_{short} \downarrow$	$E_{long} \downarrow$
Source Video	0.0532	0.262
CartoonGAN	0.0810	0.305
LSeSim	0.0691	0.272
CycleGAN	0.0607	0.240
CUT	0.0718	0.264
WhiteboxGAN	0.0670	0.266
WhiteboxGAN+Blind	0.0610	0.296
WhiteboxGAN+DVP	**0.0490**	0.2487
Baseline	0.0746	0.273
Baseline + SSA	0.0643	0.262
Baseline + PMC	0.0667	0.258
Baseline + SSA+Compound Loss	0.0517	0.250
Baseline + SSA+PMC(Ours)	0.0526	**0.237**

Table 2: Short-term and long-term warping error of different models on temporal consistency.

The qualitative results are shown in Figure 5, where generated cartoon images with the warping error heat map are illustrated for different methods. Our method preserves temporal consistency in most regions except for objects with rapid motion. Image translation methods in the first row show obvious high-frequency errors in the background. Post-processing methods decrease the warping error a lot but hurt the cartoon effect (e.g. clear edge). By contrast, our method both enhance the temporal consistency and renders great cartoon style effects.

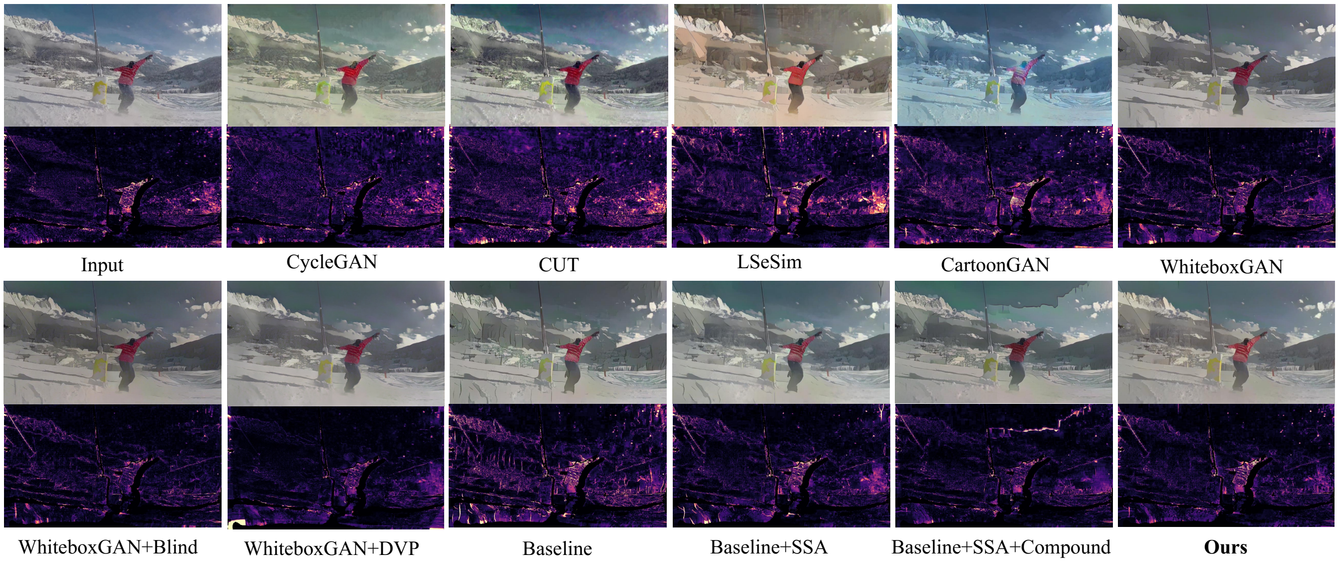

Figure 5: Qualitative comparison on temporal consistency. The first and third rows show source image and generated cartoon images of different methods. The second and fourth rows show the heat maps of warping error that indicate the differences between two adjacent video frames. Please refer to the supplementary materials for a video demonstration.

Ablation Study and Analysis

We conduct experiments on different variants of our model to evaluate the effectiveness of our proposed SSA and PMC module. We also adopt the optical flow based compound regularization (Wang et al. 2020a) for comparison with PMC. The comparison results are shown in Table 2 and Figure 5.

Effectiveness of SSA By introducing SSA, our model decreases the short-term warping error by 0.0112 and the long-term warping error by 0.011. As shown in generated images, SSA helps remove the obvious strokes generated by baseline model and better preserve the semantic-aware structure consistency. This indicates the structure consistency deriving from SSA can also benefit the temporal consistency.

Effectiveness on temporal consistency of PMC Simply applying PMC on baseline model benefits both short-term and long-term temporal consistency. By adding PMC regularization upon SSA, our model further decreases the short-term and long-term warping error by 0.0117 and 0.025. The compound regularization boosts the temporal consistency but it generates artifacts where a long edge segments the background region. By contrast, the PMC regularization can better restrict the temporal consistency without hurting the quality of generated images.

Analysis on the local patch width of SSA To explore the relationship between granularity of semantic alignment and local patch width in SSA, we conduct experiments without SSA and with SSA of different patch widths from 3 to 7. As shown in Figure 6, (1) the generated image without SSA demonstrates obvious artifacts and structure inconsistency. (2) As the increase of patch kernel size, the generator can capture more coarse semantic regions and the generated images have smoother structure.

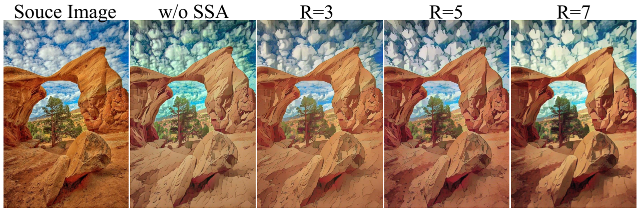

Figure 6: Images generated by our model without SSA and with SSA of different patch widths. Please zoom in to see the details.

Conclusion

In this paper, we propose a spatially-adaptive semantic alignment framework with perceptual motion consistency to generate temporal consistent cartoon videos. The proposed SSA module restores the local shift and deformation of semantic structure, which helps render semantic-aware structure consistent images. The PMC module builds a style-independent global-aware regularization on perceptual motion consistency to generate coherent cartoon videos. Experiments and ablation study demonstrate the effectiveness of our proposed modules.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grand:2018AAA0102000, National Natural Science Foundation of China: 61771457,61732007,62022083, and in part by Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2020108, CCF-Baidu Open Fund, and CAAI-Huawei MindSpore Open Fund.

References

Azulay, A.; and Weiss, Y. 2019. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20: 1–25.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017. Coherent Online Video Style Transfer. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob: 1114–1123.

Chen, J.; Liu, G.; and Chen, X. 2020. *AnimeGAN: A Novel Lightweight GAN for Photo Animation*, volume 1205 CCIS. Springer Singapore. ISBN 9789811555763.

Chen, Y.; Lai, Y. K.; and Liu, Y. J. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9465–9474.

Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; and Xu, C. 2021. Arbitrary Video Style Transfer via Multi-Channel Correlation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2): 1210–1217.

Falcon, e. a., WA. 2019. PyTorch Lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2): 167–181.

He, K.; Sun, J.; and Tang, X. 2013. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6): 1397–1409.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips): 6627–6638.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

Kim, D.; Woo, S.; Lee, J. Y.; and Kweon, I. S. 2019. Deep video inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 5785–5794.

Kingma, D. P.; and Ba, J. L. 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Lai, W. S.; Huang, J. B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M. H. 2018. Learning blind video temporal consistency. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11219 LNCS: 179–195.

Lei, C.; Xing, Y.; and Chen, Q. 2020. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS): 1–11.

Lin, C.-T.; Wu, Y.-Y.; Hsu, P.-H.; and Lai, S.-H. 2020a. Multimodal structure-consistent image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11490–11498.

Lin, J.; Wang, Y.; Chen, Z.; and He, T. 2020b. Learning to transfer: unsupervised domain translation via meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11507–11514.

Liu, M. Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips): 701–709.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Park, K.; Woo, S.; Kim, D.; Cho, D.; and Kweon, I. S. 2019. Preserving semantic and temporal consistency for unpaired video-to-video translation. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, 1248–1257.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 319–345. Springer.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Gool, L. V.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem: 724–732.

Tassano, M.; Delon, J.; and Veit, T. 2020. FastDVDNet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1351–1360.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Wang, W.; Xu, J.; Zhang, L.; Wang, Y.; and Liu, J. 2020a. Consistent video style transfer via compound regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12233–12240.

Wang, W.; Yang, S.; Xu, J.; and Liu, J. 2020b. Consistent Video Style Transfer via Relaxation and Regularization. *IEEE Transactions on Image Processing*, 29: 9125–9139.

Wang, X.; and Yu, J. 2020. Learning to Cartoonize Using White-Box Cartoon Representations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8087–8096.

Zheng, C.; Cham, T.-J.; and Cai, J. 2021. The Spatially-Correlative Loss for Various Image Translation Tasks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Zhu, J. Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob: 2242–2251.