Video-to-Video synthesis

Paper summary by Akilesh B

August 21, 2018

1 Introduction

- First work that tries to address the problem of video-to-video translation.
- Directly applying existing image to image translation works (including state-of-the-art pix2pixHD) on videos lead to output videos that are temporally inconsistent and of low perceptual quality.
- Their key contribution is well-designed generator and discriminator architectures along with spatio-temporal adversarial objective.
- Achieve high-resolution, photorealistic, temporally coherent video results on a diverse set of input formats including segmentation masks, sketches, and poses.

2 Network architecture

2.1 Notation

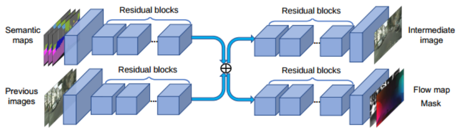
- $s_1^T \equiv \{s_1, s_2, ..., s_T\}$ denote a sequence of source images (for example image outines)
- $x_1^T \equiv \{x_1, x_2, ..., x_T\}$ be the sequence of corresponding ground truth translated images (for example colored images).
- $x_1^{\sim T} \equiv \{x_1^{\sim}, x_2^{\sim}, ..., x_T^{\sim}\}$ be the sequence of generated images.

2.2 Key idea

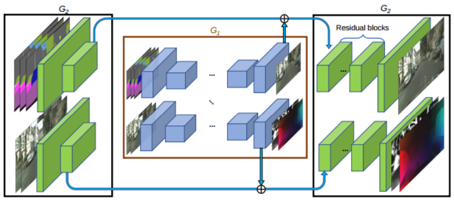
- To learn a mapping function that can convert s_1^T to $x_1^{\sim T}$ so that the conditional distribution of $x_1^{\sim T}$ given s_1^T is identical to the conditional distribution of x_1^T given s_1^T.

$$p(x_1^{\sim T}|s_1^T) = p(x_1^T|s_1^T) \tag{1}$$

- They have a coarse-to-fine generator that operate at three scales, similar to [1] as shown in 1a, 1b



(a) Generator for low resolution videos



(b) Generator for high resolution videos

Figure 1: Generator architecure

2.3 Sequential generator

Under Markovian assumption, the video frames can be generated sequentially and the generation of frame $x_t^{\sim}$ at time step t depends only on:

- Current source image(s_t)
- Past L source images(s_{t-L}^{t-1})
- Past L generated images($x_{t-L}^{\sim t-1}$)

L is typically chosen to be 2. A feed-forward network F is trained to model the conditional distribution as $x_t^{\sim} = F(x_{t-L}^{\sim t-1},\, s_{t-L}^{t})$

$$F(x_{t-L}^{\sim t-1}, s_{t-L}^{t}) = (\mathbf{1} - m_t^{\sim}) \odot w_{t-1}^{\sim}(x_{t-1}^{\sim}) + m_t^{\sim} \odot h_t^{\sim} \tag{2}$$

where $\odot$ is the element-wise product operator and $\mathbf{1}$ is an image of all ones. The first part corresponds to pixels warped from the previous frame, while the second part aims to hallucinate new pixels.

- $w_{t-1}^{\sim} = W(x_{t-L}^{\sim t-1}, s_{t-L}^{t})$ is the estimated optical flow from $x_{t-1}^{\sim}$ to $x_t^{\sim}$, and W is the optical flow prediction function.
- $h_t^{\sim} = H(x_{t-L}^{\sim t-1}, s_{t-L}^{t})$ is the hallucinated image, an image generated from scratch.
- $m_t^{\sim} = M(x_{t-L}^{\sim t-1}, s_{t-L}^{t})$ is the occlusion mask with continuous values between 0 and 1.

M, W and H are implemented using residual nets.

2.4 Conditional image discriminator D_I

D_I should output 1 for a true pair (x_t, s_t) and 0 for a fake one $(x_t^{\sim}, s_t)$.

2.5 Conditional video discriminator D_V

The purpose of D_V is to ensure that consecutive output frames resemble the temporal dynamics of a real video given the same optical flow. Suppose, w_{t-K}^{t-2} be $K-1$ optical flow for the K consecutive real images x_{t-k}^{t-1}. D_V should output 1 for a true pair $(x_{t-k}^{t-1},\, w_{t-K}^{t-2})$ and 0 for a fake one $(x_{t-k}^{\sim t-1},\, w_{t-K}^{t-2})$

2.6 Learning objective function

$$min_F\big(max_{D_I} L_I(F, D_I) + max_{D_V} L_V(F, D_V)\big) + \lambda_W L_W(F). \tag{3}$$

where L_I is the GAN loss on images defined by the conditional image discriminator D_I, L_V is the GAN loss on K consecutive frames defined by D_V, and $L_W(F)$ is the flow estimation loss as given below.

$$L_W = \frac{1}{T-1}\sum_{t=1}^{T-1}\big(\,\|\tilde{w_t} - w_t\|_1 + \|\tilde{w_t}(x_t) - x_{t+1}\|_1\,\big) \tag{4}$$

2.7 Foreground-background prior

The image hallucination function(H) is further decomposed into a foreground model $\tilde{h_{F,t}} = H_F(s_{t-L}^t)$ and a background model $\tilde{h_{B,t}} = H_B(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$.

- Optical flow can be estimated accurately in background motion and hence, background image synthesis can be generated accurately via warping.
- Optical flow estimation on a foreground object becomes difficult as it often has a large motion and occupies a small portion of the image.

Hence, the equation in 2 becomes

$$F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t) = (\mathbf{1} - \tilde{m_t}) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m_t} \odot \big((1 - m_{B,t}) \odot \tilde{h_{F,t}} + m_{B,t} \odot \tilde{h_{B,t}}\big) \tag{5}$$

3 Experiments

- Evaluate proposed approach on several datasets: Cityscapes, Apolloscape, Face video dataset, Dance video dataset etc.
- Compare their models with two baselines trained on same data:
  (a) *pix2pixHD*: State-of-the-art image to image translation work [1]
  (b) *COVST*: Based on coherent video style transfer [2] by replacing the stylization network with pix2pixHD.
- They show both subjective and objective metrics for performance evaluation using human preference scores and Frechet Inception Distance (FID).

References

[1] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J. and Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 1, No. 3, p. 5).

[2] Chen, D., Liao, J., Yuan, L., Yu, N. and Hua, G., 2017, March. Coherent online video style transfer. In Proc. Intl. Conf. Computer Vision (ICCV).