Recycle-GAN: Unsupervised Video Retargeting

Aayush Bansal[1], Shugao Ma[2], Deva Ramanan[1], and Yaser Sheikh[1,2]

[1]Carnegie Mellon University   [2]Facebook Reality Lab, Pittsburgh
http://www.cs.cmu.edu/~aayushb/Recycle-GAN/

Abstract. We introduce a data-driven approach for unsupervised video retargeting that translates content from one domain to another while preserving the style native to a domain, i.e., if contents of John Oliver's speech were to be transferred to Stephen Colbert, then the generated content/speech should be in Stephen Colbert's style. Our approach combines both spatial and temporal information along with adversarial losses for content translation and style preservation. In this work, we first study the advantages of using spatiotemporal constraints over spatial constraints for effective retargeting. We then demonstrate the proposed approach for the problems where information in both space and time matters such as face-to-face translation, flower-to-flower, wind and cloud synthesis, sunrise and sunset.

1   Introduction

We present an unsupervised data-driven approach for video retargeting that enables the transfer of sequential content from one domain to another while preserving the style of the target domain. Such a content translation and style preservation task has numerous applications including human motion and face translation from one person to other, teaching robots from human demonstration, or converting black-and-white videos to color. This work also finds application in creating visual content that is hard to capture or label in real world settings, e.g., aligning human motion and facial data of two individuals for virtual reality, or labeling night data for a self-driving car. Above all, the notion of content translation and style preservation transcends pixel-to-pixel operation, into a more semantic and abstract human understandable concepts.

Current approaches for retargeting can be broadly classified into three categories. The first set is specifically designed for domains such as human faces [5,41,42]. While these approaches work well when faces are fully visible, they fail when applied to occluded faces (virtual reality) and lack generalization to other domains. The work on paired image-to-image translation [23] attempts to generalize across domain but requires manual supervision for labeling and alignment. This requirement makes it hard for the use of such approaches as manual alignment or labeling is not possible in many domains. The third category of work attempts unsupervised and unpaired image translation [26,53]. They enforce a cyclic consistency [51] on unpaired 2D images and learn a transformation from one domain to another. However, unpaired 2D images alone are not sufficient
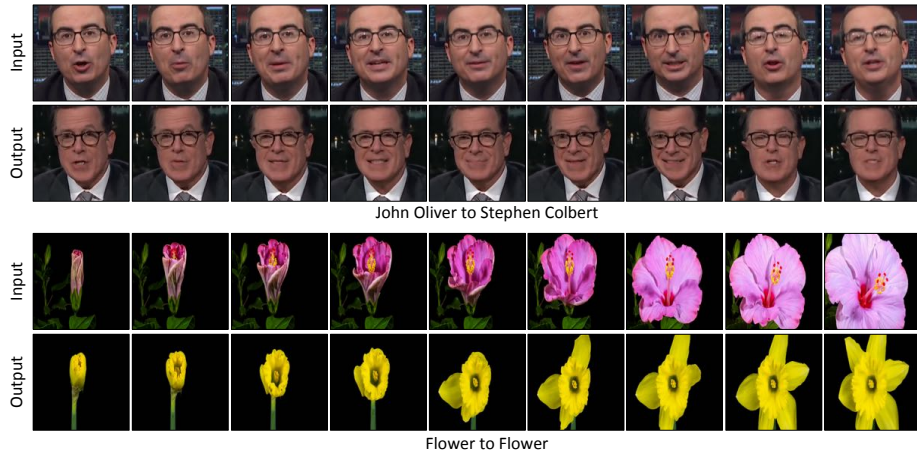
Fig. 1. Our approach for video retargeting used for faces and flowers. The top row shows translation from John Oliver to Stephen Colbert. The bottom row shows how a synthesized flower follows the blooming process with the input flower. The corresponding videos are available on the project webpage.

for video retargeting. Firstly, it is not able to pose sufficient constraints on optimization and often leads to bad local minima or a perceptual mode collapse making it hard to generate the required output in the target domain. Secondly, the use of the spatial information alone in 2D images makes it hard to learn the *style* of a particular domain as stylistic information requires temporal knowledge as well.

In this work, we make two specific observations: (i) the use of temporal information provides more constraints to the optimization for transforming one domain to other and helps in reaching a better local minima; (ii) the combined influence of spatial and temporal constraints helps in learning the style characteristic of an identity in a given domain. Importantly, temporal information is freely available in videos (available in abundance on web) and therefore no manual supervision is required. Figure 1 shows an example each of translation for human faces, and flowers. Without any manual supervision and domain-specific knowledge, our approach learns this *retargeting* from one domain to the other using publicly available video data on the web from both domains.

Our contributions : We introduce a new approach that incorporates spatiotemporal cues with conditional generative adversarial networks [15] for video retargeting. We demonstrate the advantages of spatiotemporal constraints over spatial constraints for image-to-labels, and labels-to-image in diverse environmental settings. We then present the proposed approach in learning better association between two domains, and its importance for self-supervised content alignment of the visual data. Inspired by the ever-existing nature of space-time, we qualitatively demonstrate the effectiveness of our approach for various natural processes
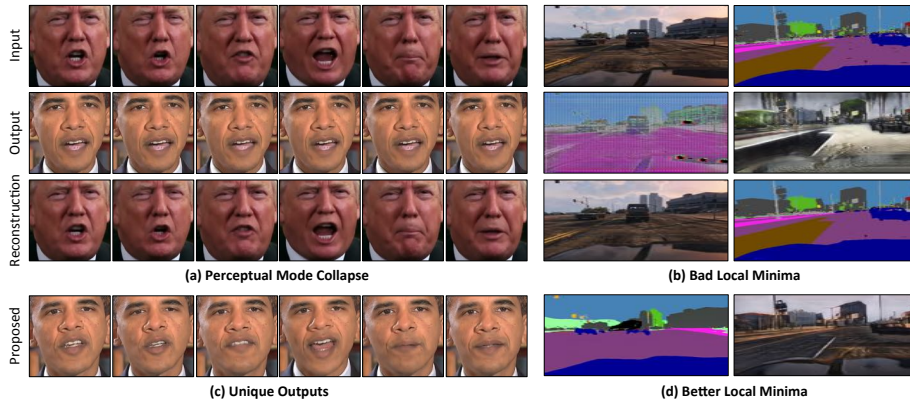
Fig. 2. Spatial cycle consistency is not sufficient: We show two examples illustrating why spatial cycle consistency alone is not sufficient for the optimization. (a) shows an example of *perceptual* mode-collapse while using Cycle-GAN [53] for Donald Trump to Barack Obama. First row shows the input of Donald Trump, and second row shows the output generated. The third row shows the output of reconstruction that takes the second row as input. The second row looks similar despite different inputs; and the third row shows output similar to first row. On a very close observation, we found that a few pixels in second row were different (but not perceptually significant) and that was sufficient to get the different reconstruction; (b) shows another example for image-to-labels and labels-to-image. While the generator is not able to generate the required output for the given input in both the cases, it is still able to perfectly reconstruct the input. Both the examples suggest that the spatial cyclic loss is not sufficient to ensure the required output in another domain because the overall optimization is focussed on reconstructing the input. However as shown in (c) and (d) , **we get better outputs with our approach combining the spatial and temporal constraints**. Videos for face comparison are available on project webpage.

such as face-to-face translation, flower-to-flower, synthesizing clouds and winds, aligning sunrise and sunset.

2   Related Work

A variety of work dealing with image-to-image translation [11,17,23,40,53] and style translation [4,10,19] exists. In fact a large body of work in computer vision and computer graphics is about an image-to-image operation. While the primary efforts were on inferencing semantic [30], geometric [1,9], or low-level cues [48], there is a renewed interest in synthesizing images using data-driven approaches by the introduction of generative adversarial networks [15]. This formulation has been used to generate images from cues such as a low-resolution image [8,28], class labels [23], and various other input priors [21,35,49]. These approaches, however, require an input-output pair to train a model. While it is feasible to label data for a few image-to-image operations, there are numerous tasks for

which it is non-trivial to generate input-output pairs for training supervision. Recently, Zhu et al. [53] propose to use the cycle-consistency constraint [51] in adversarial learning framework to deal with this problem of unpaired data, and demonstrate effective results for various tasks. Cycle-consistency [26,53] enables many image-to-image translation tasks without any expensive manual labeling. Similar ideas have also found application in learning depth cues in an unsupervised manner [14], machine translation [47], shape correspondences [20], point-wise correspondences [51,52], or domain adaptation [18].

The variants of Cycle-GAN [53] have been applied to various temporal domains [14,18]. However, they consider only the spatial information in 2D images, and ignore the temporal information for optimization. We observe two major limitations: (1). **Perceptual Mode Collapse:** there are no guarantees that cycle consistency would produce perceptually unique data to the inputs. In Figure 2, we show the outputs of a model trained for Donald Trump to Barack Obama, and an example for image-to-labels and labels-to-image. We find that for different inputs of Donald Trump, we get perceptually similar outputs of Barack Obama. We observe that these outputs have some unique encoding that enables them to reconstruct images similar to the inputs. We see similar behavior for image-to-labels and labels-to-image in Figure 2-(b); (2). **Tied Spatially to Input:** Due to the reconstruction loss on the input itself, the optimization is forced to learn a solution that is closely tied to the input. While this is reasonable for the problems where only spatial transformation matters (such as horse-to-zebra, apples-to-oranges, or paintings etc.), it is important for the problems where temporal and stylistic information is required for synthesis (prominently face-to-face translation). In this work, we propose a new formulation that utilizes both spatial and temporal constraints along with the adversarial loss to overcome these two problems. In Figure 2-(c, d), we show the outputs generated using proposed formulation that overcomes the above mentioned problems. We posit that this is due to more constraints available for an under-constrained optimization.

The use of GANs [15] and variational auto-encoder [27] have also found a way for synthesizing videos and temporal information. Walker et al. [45] use temporal information to predict future trajectories from a single image. Recent work [16,44,46] used temporal models to predict long term future poses from a single 2D image. MoCoGAN [43] decomposes motion and content to control video generation. Similarly, Temporal GAN [39] employs a temporal generator and an image generator that generates a set of latent variables and image sequences respectively. While relevant, this prior work focuses on predicting future intent from single images at test time or generating videos from a random noise. Concurrently, MoCoGAN [43] shows examples of image-to-video translation using their formulation. Different from these methods, our focus is on a general video-to-video translation where the input video can control the output in a spirit similar to image-to-image translation. To this end, we can generate high-resolution videos of arbitrary length using our approach whereas prior work [39,43] generate only 16 frames of 64×64.
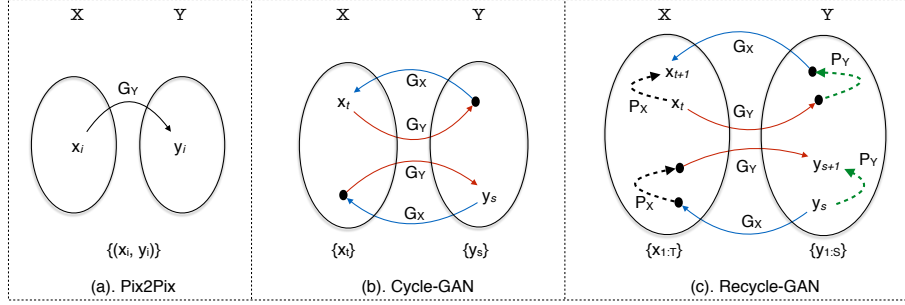
Fig. 3. We contrast our work with two prominent directions in image-to-image translation. (a) **Pix2Pix** [23]: Paired data is available. A simple function (Eq. 1) can be learnt via regression to map $X \rightarrow Y$. (b) **Cycle-GAN** [53]: The data is not paired in this setting. Zhu et al. [53] proposed to use cycle-consistency loss (Eq. 3) to deal with the problem of unpaired data. (c) **Recycle-GAN**: The approaches so far have considered independent 2D images only. Suppose we have access to unpaired but *ordered streams* $(x_1, x_2, \ldots, x_t, \ldots)$ and $(y_1, y_2 \ldots, y_s, \ldots)$. We present an approach that combines spatiotemporal constraints (Eq. 5). See Section 3 for more details.

Spatial & Temporal Constraints : Spatial and temporal information is known to be an integral sensory component that guides human action [12]. There exists a wide literature utilizing these two constraints for various computer vision tasks such as learning better object detectors [34], action recognition [13] etc. In this work, we take a first step to exploit spatiotemporal constraints for video retargeting and unpaired image-to-image translation.

Learning Association: Much of computer vision is about learning association, be it learning high-level image classification [38], object relationships [32], or point-wise correspondences [2,24,29,31]. However, there has been relatively little work on learning association for aligning the content of different videos. In this work, we use our model trained with spatiotemporal constraints to align the semantical content of two videos in a self-supervised manner, and do automatic alignment of the visual data without any additional supervision.

3   Method

Assume we wish to learn a mapping $G_Y : X \rightarrow Y$. The classic approach tunes G_Y to minimize reconstruction error on paired data samples $\{(x_i, y_i)\}$ where $x_i \in X$ and $y_i \in Y$:

$$\min_{G_Y} \sum_i ||y_i - G_Y(x_i)||^2. \tag{1}$$

Adversarial loss: Recent work [23,15] has shown that one can improve the learned mapping by tuning it with a discriminator D_Y that is adversarially

trained to distinguish between real samples of y from generated samples $G_Y(x)$:

$$\min_{G_Y} \max_{D_Y} L_g(G_Y, D_Y) = \sum_s \log D_Y(y_s) + \sum_t \log(1 - D_Y(G_Y(x_t))), \quad (2)$$

Importantly, we use a formulation that does *not* require paired data and only requires access to individual samples $\{x_t\}$ and $\{y_s\}$, where different subscripts are used to emphasize the lack of pairing.

Cycle loss: Zhu et al. [53] use cycle consistency [51] to define a reconstruction loss when the pairs are not available. Popularly known as Cycle-GAN (Fig. 3-b), the objective can be written as:

$$L_c(G_X, G_Y) = \sum_t ||x_t - G_X(G_Y(x_t))||^2. \quad (3)$$

Recurrent loss: We have so far considered the setting when static data is available. Instead, assume that we have access to unpaired but *ordered streams* $(x_1, x_2, \ldots, x_t, \ldots)$ and $(y_1, y_2 \ldots, y_s, \ldots)$. Our motivating application is learning a mapping between two videos from different domains. One option is to ignore the stream indices, and treat the data as an unpaired *and unordered* collection of samples from X and Y (e.g., learn mappings between shuffled video frames). We demonstrate that much better mapping can be learnt by taking advantage of the temporal ordering. To describe our approach, we first introduce a recurrent temporal predictor P_X that is trained to predict future samples in a stream given its past:

$$L_\tau(P_X) = \sum_t ||x_{t+1} - P_X(x_{1:t})||^2, \quad (4)$$

where we write $x_{1:t} = (x_1 \ldots x_t)$.

Recycle loss: We use this temporal prediction model to define a new cycle loss across domains and *time* (Fig. 3-c) which we refer as a recycle loss:

$$L_r(G_X, G_Y, P_Y) = \sum_t ||x_{t+1} - G_X(P_Y(G_Y(x_{1:t})))||^2, \quad (5)$$

where $G_Y(x_{1:t}) = (G_Y(x_1), \ldots, G_Y(x_t))$. Intuitively, the above loss requires *sequences* of frames to map back to themselves. We demonstrate that this is a much richer constraint when learning from unpaired data streams in Figure 4.

Recycle-GAN: We now combine the recurrent loss, recycle loss, and adversarial loss into our final Recycle-GAN formulation:

$$\min_{G,P} \max_D L_{rg}(G, P, D) = L_g(G_X, D_X) + L_g(G_Y, D_Y) +$$

$$\lambda_{rx} L_r(G_X, G_Y, P_Y) + \lambda_{ry} L_r(G_Y, G_X, P_X) + \lambda_{\tau x} L_\tau(P_X) + \lambda_{\tau y} L_\tau(P_Y).$$

Inference: At test time, given an input video with frames $\{x_t\}$, we would like to generate an output video. The simplest strategy would be directly using the

| Input | Cycle-GAN | Recycle-GAN | Ground Truth | Input | Cycle-GAN | Recycle-GAN | Ground Truth |
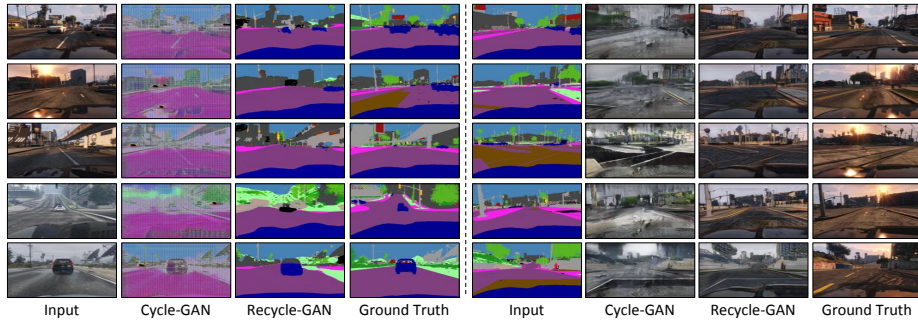
Fig. 4. We compare the performance of our approach for image2labels and labels2image with Cycle-GAN [53] on a held out data of Viper dataset [36] for various environmental conditions.

trained G_Y to generate a video frame-by-frame $y_t = G_Y(x_t)$. Alternatively, one could use the temporal predictor P_Y to smooth the output:

$$y_t = \frac{G_Y(x_t) + P_Y(G_Y(x_{1:t-1}))}{2},$$

where the linear combination could be replaced with a nonlinear function, possibly learned with the original objective function. However, for simplicity, we produce an output video by simple single-frame generation. This allows our framework to be applied to both videos and single images at test-time, and produces fairer comparison to spatial approach.

Implementation Details: We adopt much of the training details from Cycle-GAN [53] to train our spatial translation model, and Pix2Pix [23] for our temporal prediction model. The generative network consists of two convolution (downsampling with stride-2), six residual blocks, and finally two upsampling convolution (each with a stride 0.5). We use the same network architecture for G_X, and G_Y. The resolution of the images for all the experiments is set to 256×256. The discriminator network is a 70×70 PatchGAN [23,53] that is used to classify a 70×70 image patch if it is real or fake. We set all $\lambda_s = 10$. To implement our temporal predictors P_X and P_Y, we concatenate the last two frames as input to a network whose architecture is identical to U-Net architecture [23,37].

4   Experiments

We now study the influence of spatiotemporal constraints over spatial cyclic constraints. Because our key technical contribution is the introduction of temporal constraints in learning unpaired image mappings, the natural baseline is Cycle-GAN [53], a widely adopted approach for exploiting spatial cyclic consistency alone for an unpaired image translation. We first present quantitative results on domains where ground-truth correspondence between input and output videos

are known (e.g., a video where each frame is paired with a semantic label map). Importantly, this correspondence pairing is *not available* to either Cycle-GAN or Recycle-GAN, but used only for evaluation. We then present qualitative results on a diverse set of videos with unknown correspondences, including video translations across different human faces and temporally-intricate events found in nature (flowers blooming, sunrise/sunset, time-lapsed weather progressions).

4.1   Quantitative Analysis

We use publicly available Viper [36] dataset for image-to-labels and labels-to-image to evaluate our findings. This dataset is collected using computer games with varying realistic content and provides densely annotated pixel-level labels. Out of the 77 different video sequences consisting of diverse environmental conditions, we use 57 sequences for training our model and baseline. The held-out 20 sequences are used for evaluation. The goal for this evaluation is not to achieve the state-of-the-art performance but to compare and understand the advantage of spatiotemporal cyclic consistency over the spatial cyclic consistency [53].

Criterion	Approach	day	sunset	rain	snow	night	all
MP	Cycle-GAN	35.8	38.9	51.2	31.8	27.4	35.5
	Recycle-GAN	**48.7**	**71.0**	**60.9**	57.1	**45.2**	**56.0**
	Combined	**48.7**	70.0	60.1	**58.9**	33.7	53.7
AC	Cycle-GAN	7.8	6.7	7.4	7.0	4.7	7.1
	Recycle-GAN	11.9	12.2	**10.5**	11.1	**6.5**	11.3
	Combined	**12.6**	**13.2**	10.1	**13.3**	5.9	**12.4**
IoU	Cycle-GAN	4.9	3.9	4.9	4.0	2.2	4.2
	Recycle-GAN	7.9	9.6	7.1	8.2	**4.1**	8.2
	Combined	**8.5**	**13.2**	**10.1**	**9.6**	3.1	**8.9**

Table 1. Image-to-Labels (Semantic Segmentation): We use the Viper [36] dataset to evaluate the performance improvement when using spatiotemporal constraints as opposed to only spatial cyclic consistency [53]. We report results using three criteria: (1). Mean Pixel Accuracy (**MP**); (2). Average Class Accuracy (**AC**); and (3). Intersection over union (**IoU**). We observe that our approach achieves better performance than prior work, and combining both leads to further better performance.

While the prior works [23,53] have mostly used Cityscapes dataset [7], we could not use it for our evaluation. Primarily the labelled images in Cityscapes are not continuous video sequences. That is consecutive frame pairs are quite different from other another. As such it is not trivial to use a temporal predictor. We used Viper as a proxy for Cityscapes because the task is similar and that dataset contains dense video annotations. Additionally, a concurrent work [3] on

unsupervised video-to-video translation also uses Viper dataset for evaluation. However, their experiments are restricted to a small subset of sequences from daylight category. In this work, we use all the environmental conditions available in the dataset.

Image-to-Labels : In this setting, we input a RGB image to the generator that output segmentation label maps. We compute three metrics to compare the output ofs two approaches: (1). Mean Pixel Accuracy (**MP**); (2). Average Class Accuracy (**AC**); (3). Intersection over Union (**IoU**). These statistics are computed using the ground truth for the held-out sequences from different environmental conditions. Table 1 contrast the performance of our approach (Recycle-GAN) with Cycle-GAN. We observe that Recycle-GAN achieves better performance than Cycle-GAN, and combining both losses further improved it.

Labels-to-Image : In this setting, we input a segmentation label map to generator and output an image that is close to a real image. The goal of this evaluation is to compare the quality of output images obtained from both approaches. We follow Pix2Pix [23] for this evaluation. We use the generated images from each of the algorithm with a pre-trained FCN model. We then compute the performance of synthesized images against the real images to compute a normalized FCN-score. Higher performance on this criterion suggest that generated image is closer to the real images. Table 2 compares the performance of our approach with Cycle-GAN. We observe that Recycle-loss does better than Cycle-loss, and combining both the losses led to significantly better outputs. Figure 4 qualitatively compares our approach with Cycle-GAN.

Approach	day	sunset	rain	snow	night	all
Cycle-GAN	0.33	0.27	0.39	0.29	0.37	0.30
Recycle-GAN	0.33	0.51	0.37	0.43	0.40	0.39
Combined	**0.42**	**0.61**	**0.45**	**0.54**	**0.49**	**0.48**

Table 2. Normalized FCN score for Labels-to-Image: We use a pre-trained FCN-style model to evaluate the quality of synthesized images over real images using the Viper [36] dataset. Higher performance on this criteria suggest that the output of a particular approach produces images that look closer to the real images.

In these experiments, we make two observations: (i) Cycle-GAN learns a good translation model within a few initial iterations (seeing only a few examples) but this model degraded as reconstruction loss started to decrease. We believe that minimizing reconstruction loss alone on input led it to a bad local minima. Our formulation provided more constraints and led it to a better local minima; (ii) Cycle-GAN learns a better translation model for Cityscapes as opposed to Viper. Cityscapes consists of images from mostly daylight and agreeable weather. This is not the case with Viper as it is rendered. It has a large and varied distribution of different illumination (day, night) and weather conditions (snow, rain). This

Input

Output

John Oliver to Stephen Colbert   Stephen Colbert to John Oliver

Input

Output

John Oliver to a Cartoon Character   Barack Obama to Donald Trump   MLK to Barack Obama

Fig. 5. Face to Face: The top row shows multiple examples of face-to-face between John Oliver and Stephen Colbert using our approach. The bottom row shows example of translation from John Oliver to a cartoon character, Barack Obama to Donald Trump, and Martin Luther King Jr. (MLK) to Barack Obama. Without any input alignment or manual supervision, our approach could capture stylistic expressions for these public figures. As an example, John Oliver's dimple while smiling, the shape of mouth characteristic of Donald Trump, and the facial mouth lines and smile of Stephen Colbert. More results and videos are available on our project webpage.

makes it harder to learn a good mapping because for each labelled input as there are potentially many output images. We find that standard conditional GANs suffer from mode collapse in such scenarios and produce "average" outputs (as pointed by prior works [2]). Our experiments suggest that having more constraints help ameliorate such challenging translation problems.

4.2   Qualitative Analysis

Face to Face: We use publicly available videos of various public figures for face to face translation. The faces are extracted using facial keypoints generated using the OpenPose Library[6] and are manually curated to remove false positives. Figure 5 shows an example of face-to-face translation between John Oliver and Stephen Colbert, Barack Obama to Donald Trump, and Martin Luther King Jr. (MLK) to Barack Obama, and John Oliver to a cartoon character. Note that without any supervisory signal or manual alignment, our approach learns face-to-face translation and captures stylistic expression for these personalities, such as the dimple on the face of John Oliver while smiling, the characteristic shape of mouth of Donald Trump, and the mouth lines for Stephen Colbert.
Flower to Flower: Extending from faces and other traditional translations, we demonstrate our approach for flowers. We extracted the time-lapse of various flowers from publicly available videos. The time-lapses show the blooming of

Fig. 6. Flower to Flower: We shows two examples of flower-to-flower translation. Note the smooth transition from Left to Right. These results can be best visualized using videos on our project webpage.
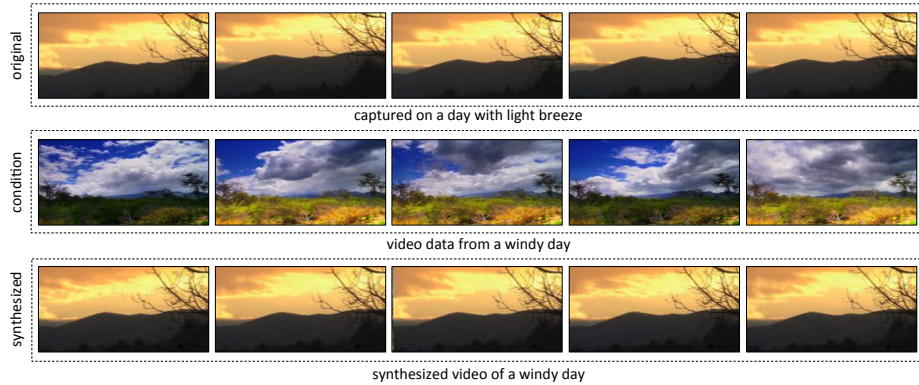
different flowers but without any sync. We use our approach to align the content, i.e. both flowers bloom or die together. Figure 6 shows the output of approach to learn association between the events of different flowers' life.

4.3   Video Manipulation via Retargeting

Clouds & Wind Synthesis: Our approach can be used to synthesize a new video that has the required weather condition such as clouds and wind without the need for recapturing. We use the given video, and video data from target environmental condition as two domains in our experiment. The conditional video, and trained translation model is then used to generate the required output.

   We collected the video data for various wind and cloud conditions, such as calm day or windy day for this experiment. We convert a calm-day to a windy-day, and a windy-day to a calm-day using our approach, without modifying the aesthetics of the place. Figure 7 shows an example of synthesizing clouds and winds on a windy day at a place when the only information available was a video captured at same place with a light breeze. Additional videos for these clouds and wind synthesis are available on our project webpage.

Sunrise & Sunset: We extracted the sunrise and sunset data from various web videos, and show how our approach could be used for both video manipulation and content alignment. This is similar to our experiments on clouds and wind synthesis. Figure 8 shows an example of synthesizing a sunrise video from an original sunset video by conditioning it on a random sunrise video. We also show examples of alignment of various sunrise and sunset scenes.

original

captured on a day with light breeze

condition

video data from a windy day

synthesized

synthesized video of a windy day

Fig. 7. Synthesizing Clouds & Winds: We use our approach to synthesize clouds and winds. The top row shows example frames of a video captured on a day with light breeze. We condition it on video data from a windy data (shown in second row) by learning a transformation between two domains using our approach. The last row shows the output synthesized video with the clouds and trees moving faster (giving a notion of wind blowing). Refer to the videos on our project webpage for better visualization and more examples.

Note: We refer the reader to our project webpage for different videos synthesized using our approach. We have also added the results using both Cycle-loss and Recycle-loss on our project webpage.

4.4   Human Studies

We performed human studies on the synthesized output, particularly faces and flowers, following the protocol of MoCoGAN [43] who also evaluated videos. However, our analysis consist of three parts: (1). We show the synthesized videos individually from both Cycle-GAN and ours to 15 sequestered human subjects, and asked them if it is a real video or a generated video. The subjects misclassified 28.3% times generated videos from our approach as real, and 7.3% times for Cycle-GAN. (2). We show the synthesized videos from Cycle-GAN and our approach simultaneously, and asked them to tell which one looks more natural and realistic. Human subjects chose the videos synthesized from our approach 76% times, 8% times Cycle-GAN, and 16% times they were confused. (3). We show the video-to-video translation. This is an extension of (2), except now we also include input and ask which translation looks more realistic and natural. We showed each video to 15 human subjects. The human subjects selected our approach 74.7% times, 13.3% times they selected Cycle-GAN, and 12% times they were confused. From the human study, we can clearly see that combining spatial and temporal constraints lead to better retargeting.
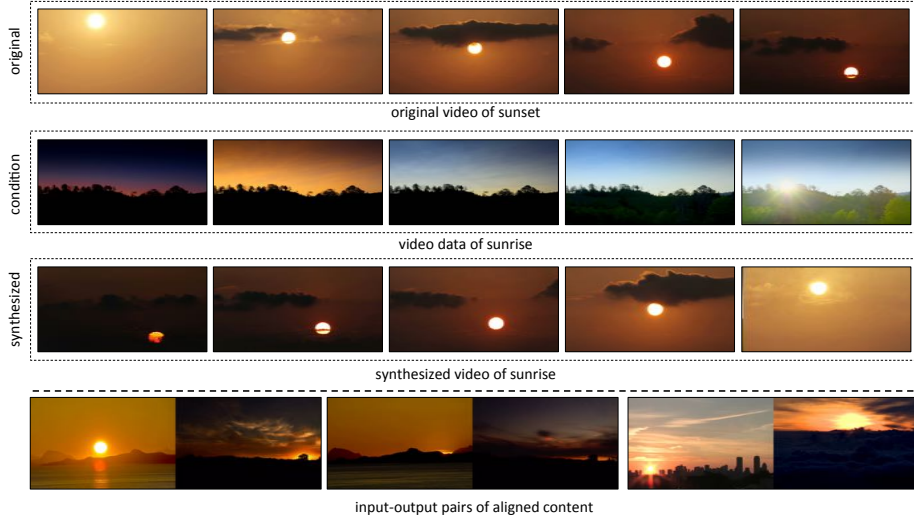
Fig. 8. Sunrise & Sunset: We use our approach to manipulate and align the videos of sunrise and sunset. The top row shows example frames from a sunset video. We condition it on video data of sunrise (shown in second row) by learning a transformation between two domains using our approach. The third row shows example frames of new synthesized video of sunrise. Finally, the last row shows random examples of input-output pair from different sunrise and sunset videos. Videos and more examples are available on our project webpage.

4.5 Failure Example: Learning association beyond data distribution

We show an example of transformation from a real bird to a origami bird to demonstrate a case where our approach failed to learn the association. The real bird data was extracted using web videos, and we used the origami bird from the synthesis of Kholgade et al. [25]. Figure 9 shows the synthesis of origami bird conditioned on the real bird. While the real bird is sitting, the origami bird stays and attempts to imitate the actions of real bird. The problem comes when the bird begins to fly. The initial frames when the bird starts to fly are fine. After some time the origami bird reappears. From an association perspective, the origami bird should not have reappeared. Looking back at the training data, we found that the original origami bird data does not have a example of frame without the origami bird, and therefore our approach is not able to associate an example when the real bird is no longer visible. Perhaps, our approach could only learn to interpolate over a given data distribution and fails to capture anything beyond it. A possible way to address this problem is by using a lot of training data so that it encapsulates all possible scenarios for an effective interpolation.

Fig. 9. Failure Example: We present the failure in association/synthesis for our approach using a transformation from a *real* bird to an *origami* bird. While the origami bird (output) is trying to imitate the real bird (input) when it is sitting (Column 1 - 4), and also flies away when the real bird flies (Column 5 - 6). We observe that it reappears after sometime (red bounding box in Column 7) in a flying mode while the real bird didn't exist in the input. Our algorithm is not able to make transition of association when the real bird is completely invisible, and so it generated a random flying origami.

5   Discussion & Future Work

In this work, we explore the influence of spatiotemporal constraints in learning video retargeting and image translation. Unpaired video/image translation is a challenging task because it is unsupervised, and lacks any correspondences between training samples from the input and output space. We point out that many natural visual signals are inherently spatiotemporal in nature, which provides strong temporal constraints for free. This results in significantly better mappings. We also point out that unpaired and unsupervised video retargeting and image translation is an under-constrained problem. More constraints using auxiliary tasks from the visual data itself (as used for other vision tasks [33,50]) could help in learning better transformation models.

Recycle-GANs learn both a mapping function and a recurrent temporal predictor. Thus far, our results make use of only the mapping function, so as to facilitate fair comparisons with previous work. But it is natural to synthesize target videos by making use of both the single-image translation model and the temporal predictor. Additionally, the notion of style in video retargeting can be incorporated more precisely by using spatiotemporal generative models as this would allow to even learn the speed of generated output. E.g. Two people may have different ways of content delivery and that one person can take longer than other to say the same thing. A true notion of style should be able to generate even this variation in time required for delivering speech/content. We believe that better spatiotemporal neural network architecture could attempt this problem in near future. Finally, our work could also utilize the concurrent approach from Huang et al. [22] to learn a one-to-many translation model.

References

1. Bansal, A., Russell, B., Gupta, A.: Marr Revisited: 2D-3D model alignment via surface normal prediction. In: CVPR (2016)
2. Bansal, A., Sheikh, Y., Ramanan, D.: PixelNN: Example-based image synthesis. In: ICLR (2018)
3. Bashkirova, D., Usman, B., Saenko, K.: Unsupervised video-to-video translation. CoRR **abs/1806.03698** (2018)
4. Brand, M., Hertzmann, A.: Style machines. ACM Trans. Graph. (2000)
5. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (2014)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
7. 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: CVPR (2016)
8. Denton, E.L., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a laplacian pyramid of adversarial networks. In: NIPS (2015)
9. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
10. Freeman, W.T., Tenenbaum, J.B.: Learning bilinear models for two-factor problems in vision. In: CVPR (1997)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
12. Gibson, J.J.: The ecological approach to visual perception. (1979)
13. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: ActionVLAD: Learning spatio-temporal aggregation for action classification. In: CVPR (2017)
14. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. In: NIPS (2014)
16. He, J., Lehrmann, A., Marino, J., Mori, G., Sigal, L.: Probabilistic video generation using holistic attribute control. In: ECCV (2018)
17. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. ACM Trans. Graph. (2001)
18. Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
19. Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. ACM Trans. Graph. (2005)
20. Huang, Q.X., Guibas, L.: Consistent shape maps via semidefinite programming. In: Eurographics Symposium on Geometry Processing (2013)
21. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J.E., Belongie, S.J.: Stacked generative adversarial networks. In: CVPR (2017)
22. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
24. Kanazawa, A., Jacobs, D.W., Chandraker, M.: Warpnet: Weakly supervised matching for single-view reconstruction. In: CVPR (2016)

25. Kholgade, N., Simon, T., Efros, A., Sheikh, Y.: 3d object manipulation in a single photograph using stock 3d models. ACM Trans. Graph. (2014)
26. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
29. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. (2011)
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional models for semantic segmentation. In: CVPR (2015)
31. Long, J., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: NIPS (2014)
32. Malisiewicz, T., Efros, A.A.: Beyond categories: The visual memex model for reasoning about object relationships. In: NIPS (2009)
33. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: AAAI (2018)
34. Misra, I., Shrivastava, A., Hebert, M.: Watch and learn: Semi-supervised learning of object detectors from videos. In: CVPR (2015)
35. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015)
36. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: International Conference on Computer Vision (ICCV) (2017)
37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV (2015)
39. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: ICCV (2017)
40. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
41. Thies, J., Zollhofer, M., Niessner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. (2015)
42. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
43. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR (2018)
44. Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H.: Learning to generate long-term future via hierarchical prediction. In: ICML (2017)
45. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from variational autoencoders. In: ECCV (2016)
46. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017)
47. Xia, Y., He, D., Qin, T., Wang, L., Yu, N., Liu, T., Ma, W.: Dual learning for machine translation. In: NIPS (2016)
48. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)

49. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
50. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
51. Zhou, T., Krähenbühl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: CVPR (2016)
52. Zhou, T., Lee, Y.J., Yu, S.X., Efros, A.A.: FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In: CVPR (2015)
53. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)