Unsupervised Multimodal Video-to-Video Translation via Self-Supervised Learning

Kangning Liu[1,2*] Shuhang Gu[2*] Andrés Romero[2]
Radu Timofte[2]
[1]Center for Data Science, New York University, USA [2]Computer Vision Lab, ETH Zürich, Switzerland

[*]Equal contributions

In this supplementary material, we provide more details and results of our method as follows:

1. Additional loss details and implementation details;
2. More experimental results with the resolution of 256×256;
   - Translation between image and label: comparison with the baseline, multimodal results and a 1680-frame style-consistent sequence;
   - Translation on other datasets: Rain and Snow, Sunset and Day, Viper and Cityscapes;
3. More experimental results with the resolution of 128×128;
4. Additional ablation experiments details.

Due to the size constraint, we did not include all the video material in the supplementary material.

1. Additional Loss Details and Implementation Details

1.1. Loss functions for the discriminator

In this section, we provide more details of our image-level (D^{img}), video-level (D^{vid}), and style latent (D_Z) discriminator losses. For the purpose of simplicity, we only present the loss functions for domain A, and the loss functions for domain B are defined following the same set of equations. Our adversarial loss is based on Relativistic GAN (RGAN) [7], which tries to predict the probability that a real sample is relatively more realistic than a fake one.

Image level discriminator loss The loss term D_A^{img} is defined as follows:

$$\begin{aligned} L_{D_A^{img}}^{GAN} = \frac{1}{2(T-2)} \sum_{i=2}^{i=T-1} [D_A^{img}(a_i) - D_A^{img}(a_i^{interp}) - 1]^2 \\ + \frac{1}{2T} \sum_{i=1}^{i=T} [D_A^{img}(a_i) - D_A^{img}(a_i^{trans}) - 1]^2. \end{aligned} \tag{1}$$

Video level discriminator loss D_A^{vid} for domain A is defined as follows:

$$L_{D_A^{vid}}^{GAN} = [D_A^{vid}(a_{1:T}) - D_A^{vid}(a_{1:T}^{trans}) - 1]^2. \tag{2}$$

Style latent variable discriminator loss This loss term (D_{Z_A}) for the style domain A is defined as follows:

$$L_{D_{Z_A}}^{GAN} = [D_{Z_A}(z_a^{prior}) - D_{Z_A}(z_a^{post}) - 1]^2. \tag{3}$$

1.2. Network structure

Style Encoder, Content Encoder and Content Decoder Our style encoder is similar to the one used in Augment CycleGAN [1]. Under the shared content space assumption [8], we decompose the style-conditioned Resnet-Generator used in Augment CycleGAN [1] into a Content Encoder and a Content Decoder. Moreover, when the sub-domain information is available, we assign part of the style latent variable to record such prior information. Concretely, we use one-hot vector to encode the sub-domain information.

RNN - Trajectory Gated Recurrent Units (TrajGRUs) Traditional RNN (Recurrent Neural Network) is based on the fully connected layer, which has limited capacity of profiting from the underlying spatio-temporal information in video sequence. In order to take full advantage of the spatial and temporal correlations, UVIT utilizes a convolutional RNN architecture in the generator. TrajGRU [10] is one variant of Convolutional RNN (Recurrent Neural Network) [12], which can actively learn the location-variant structure in the video data. It uses the input and hidden state to generate the local neighborhood set for each location at each time, thus warping the previous state to compensate for the motion information. We take two TrajGRUs to propagate the inter-frame information in both directions in the shared content space.

Discriminators (D^{img}, D^{vid}, D_Z) For the image-level discriminators D^{img}, the architecture is based on the PatchGANs [6] approach. Likewise, Video-level discriminators D^{vid} are similar to PatchGANs, yet we employ 3D convolutional filters. For the style latent variable discriminators D_Z, we use the same architecture as in Augmented CycleGAN [1].

1.3. Datasets

We validate our method using two common yet challenging datasets: Viper [9], and Cityscapes [5] datasets.

Viper has semantic label videos and scene image videos. There are 5 subdomains for the scene videos: day, sunset, rain, snow and night. The large diversity of scene scenarios makes this dataset a very challenging testing bed for the unsupervised V2V task. We quantitatively evaluate translation performance by different methods on the image-to-label and the label-to-image mapping tasks. We further conduct the translation between different subdomains of the scene videos for qualitative analysis.

Cityscapes has real-world street scene videos. As there is not subdomain information for Cityscapes, we conduct experiments without subdomain label for Cityscapes. We conduct qualitative analysis on the translation between scene videos of Cityscapes and Viper dataset. Note that there is no ground truth semantic labels for the continuous Cityscapes video sequences. The semantic labels are only available to a limited portion of none-continuous individual images. Therefore, we could not use it for our evaluation of image-to-label (semantic segmentation) performance.

1.4. Implementation Details

We train our model using images of 128×128 and 10 frames per batch in a single NVIDIA P100 GPU for the main experiments to capture temporal information with more frames. Setting the batch size to one, it takes about one week to train. Note that it takes roughly 4 days to train using 6 frames per batch.

During inference, we use video sequences of 30 frames. These 30 frames are divided into 4 smaller sequences of 10 frames with overlap. They all share the same style code to be style consistent. To get a higher resolution and show more details within the existing GPU resource constraint, we also train our model using images of 256×256 and 4 frames per batch.

The λ parameters. Video interpolation loss weight λ_{interp} is set to 10. Cycle consistency loss weight λ_{cycle} is set to 10. Style reconstruction loss weight λ_{rec} is set to 0.025.

1.5. Human Preference Score

We have conducted human subjective experiments to evaluate the visual quality of synthesized videos using the Amazon Mechanical Turk (AMT) platform.

For the video-level evaluation, we show two videos (synthesized by two different models) to AMT participants, and ask them to select which one looks more realistic regarding a video-consistency and video quality criteria.

- **UVIT (ours) / 3DCycleGAN:** Since 3DcycleGAN [3] generates consistent output with 8 frames in the original paper setting, UVIT results are organized to 8 frames for a fair comparison.
- **UVIT (ours) / Improved ReCycleGAN:** When comparing with improved RecycleGAN [2], we take each video clip with 30 frames.
- **UVIT (ours) / vid2vid:** When comparing with vid2vid [11], we take each video clip with 28 frames, following the setting in vid2vid [11].

For the image-level evaluation, we show to AMT participants two generated frames synthesized by two different algorithms, and ask them which one looks more real in visual quality.

These evaluations have been conducted for 100 videos and frame samples to assess the image-level and video-level qualities, respectively. We gathered answers from 10 different workers for each sample.

2. Higher resolution results 256 * 256

To get a higher resolution and show more details within the existing GPU resource constraint, we also train our model using images of 256×256 and 4 frames per batch. During the test time, we divide a longer sequence into sub-sequences of 4 frames with overlap. All the results of this section are trained using images of 256×256 and 4 frames per batch. Note that all visual examples in this paper are reshaped to the aspect ratio of the raw Viper image for better visual presentation.

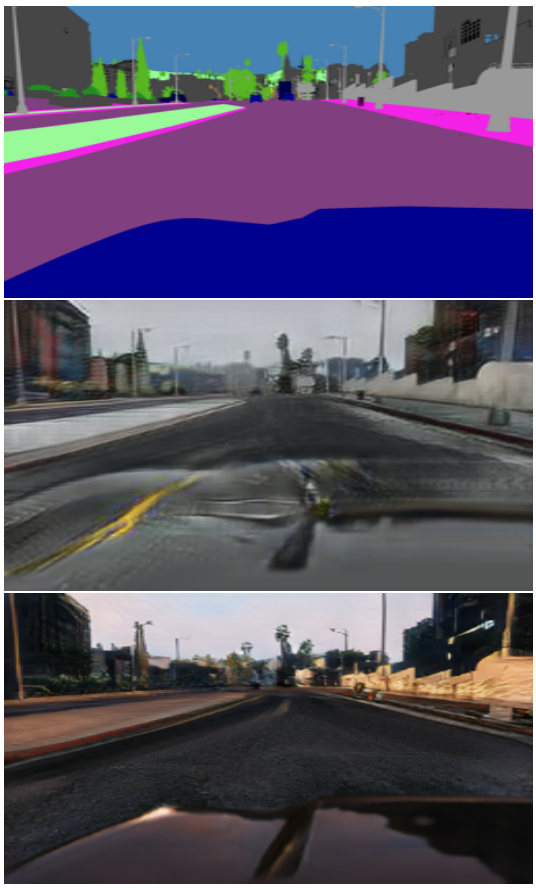

Figure 1. **Video screenshot of the video corresponding to Fig. 1 in the main paper.** From Top to Bottom: input, RecycleGAN output, UVIT output. The video is attached as $1\_HRcomapare.mp4$

2.1. Additional examples of the label-to-image qualitative comparison

In Figure 1 (corresponding video $1\_HRcompare.mp4$) and Figure 2 (corresponding video $2\_HRcomapare2.mp4$), we provide the visual examples of how our UVIT method compares with respect to RecycleGAN [2]. The RecycleGAN outputs are generated by the original code provided by the author of RecycleGAN in 256×256. Besides the video-level quality comparison from videos, we encourage the reader to also check the frame-level quality from images since *.mp4* format may fail to preserve some image-level quality.

2.2. Quantitative comparison of the label-to-image and image-to-label

In Table 1 and 2, we show quantitative results for our proposed method trained with a resolution of 256×256 and 4 frames per batch.

2.3. Label-to-image multi-subdomain and multimodality results

Video results of UVIT on label sequences to image sequences with multi-subdomain and multimodality are shown in Figure 3 and the enclosed video *3_Multimodality.mp4*. The videos are all with a length of 220 frames.

2.4. Long video example (1680 frames)

In Figure 4 and attached video $4\_long\_consistency.mp4$, we provide a video sequence example with more than 1680 frames to give a qualitative example of how our UVIT model performs in terms of style consistency. Note that the semantic labels in Viper [9] are automatic generated, however, we observe that there may still exist a little small flips in the input semantic label sequence occasionally.

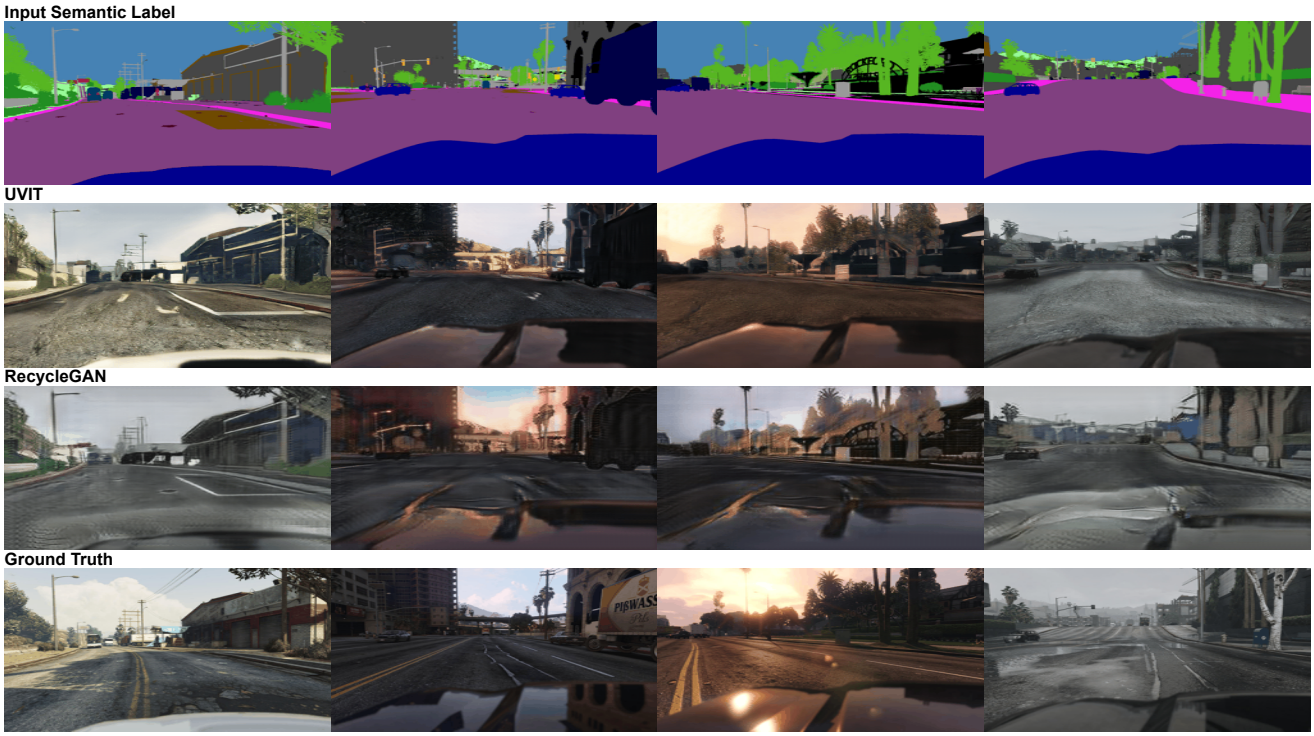

Figure 2. **Video screenshot of the comparison with RecycleGAN [2]:** We aim to compare the content consistency and image-level quality. Here the RecycleGAN results are produced by the original RecycleGAN code in a resolution of 256×256. Since there is no guarantee of style consistency for RecycleGAN, we select some RecycleGAN visual results in a small sequence length of 30 frames where style is almost consistent to compare with UVIT (ours). The corresponding video is attached as $2\_HRcomapare2.mp4$

Table 1. **Quantitative comparison between UVIT and baseline approaches on the image-to-label (Semantic segmentation) task.**(256×256 **with 4 frames per batch during training)** .Our translator effectively leverage the temporal information directly, thus producing more semantic persevering translation outcomes

Criterion	Model	Day	Sunset	Rain	Snow	Night	All
mIoU↑	ReCycleGAN (Reproduced)	10.32	11.19	11.25	9.83	7.73	10.12
	UVIT (Ours) (frame 4)	**12.05**	**12.23**	**13.37**	**11.54**	**10.49**	**11.93**
AC↑	ReCycleGAN (Reproduced)	15.80	15.79	15.93	15.57	11.47	14.85
	UVIT (Ours) (frame 4)	**17.21**	**17.41**	**18.16**	**17.37**	**14.30**	**16.50**
PA↑	ReCycleGAN (Reproduced)	54.70	55.92	57.71	50.85	49.11	53.66
	UVIT (Ours) (frame 4)	**63.44**	**61.98**	**64.72**	**60.83**	**62.05**	**62.35**

Table 2. **Quantitative comparison between UVIT and baseline approaches on the label-to-image task** (256×256 **with 4 frames per batch during training**). Better FID indicates that our translation has better visual quality and temporal consistency. We use the pre-trained network (I3D [4]) to extract features from 30-frame sequences just as the experiments in the main paper.

Criterion	Model	Day	Sunset	Rain	Snow	Night
FID↓	ReCycleGAN [2]	23.60	24.45	28.54	31.58	35.74
	UVIT (ours) (frame 4)	**18.68**	**16.70**	**20.20**	**18.27**	**19.29**

Figure 3. **Video screenshot of the label-to-image multi-subdomain and multimodality results**. Better depicted in *3_Multimodality.mp4*.

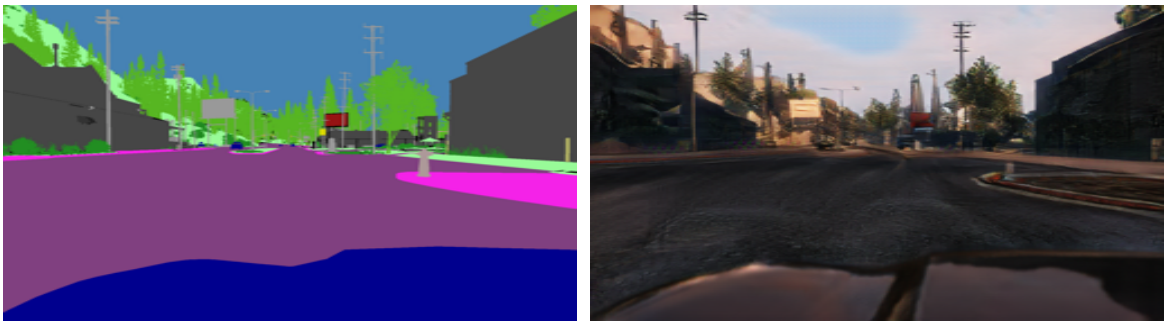

Figure 4. **Screenshot of a long style consistent translation video visual example (1680 frames)**. Left: input semantic labels; Right: UVIT translated video in sunset scenario. All frames within the video share the same style code to keep style consistency. The video is attached as $4\_long\_consistency.mp4$

Figure 5. **Screenshot of Viper Rain-and-Snow translation.** First row: real rain inputs; Second row: UVIT translated snow videos; Third row: real snow inputs; Fourth row: UVIT translated rain videos. Video is attached as *5_Rainandsnow.mp4*

2.5. Translation on other datasets

In Figure 5 and in the attached video *5_Rainandsnow.mp4*, we provide visual examples of UVIT video translation between Rain and Snow scenes in the Viper dataset. In Figure 6 and in the attached video *6_Sunsetanday.mp4*, we provide visual examples of UVIT video translation between Sunset and Day scenes in the Viper dataset. In Figure 7 and in the attached video *7_Cityscapesandviper.mp4*, we provide visual examples of UVIT video translation between Cityscapes dataset and Viper dataset. Besides the video-level quality evaluation from videos, we encourage the reader to also check the frame-level quality from images since *.mp4* format may fail to preserve some image-level quality.

Figure 6. **Screenshot of Viper Sunset-and-Day translation.** First row: real sunset inputs; Second row: UVIT translated day videos; Third row: real day inputs; Fourth row: UVIT translated sunset videos. Video is attached as *6_Sunsetandday.mp4*

Figure 7. **Screenshot of Cityscapes-and-Viper translation.** Top left: real Cityscapes input; Top right: UVIT translated Viper videos with different style codes; Second row: More UVIT translated Viper videos; Bottom left: real Viper input; Bottom right: UVIT translated Cityscapes videos with different style codes. Since the general distribution between Cityscapes and Viper may be different (*e.g.* there are more buildings in Cityscapes), the translated Viper video may differ from input Cityscapes video in class distribution to fool the discriminator, so as to be close to the class distribution in the target domain. Video is attached as *7_Cityscapesandviper.mp4*

3. Additional examples of the label-to-image qualitative comparison 128 * 128

Note that all visual examples in this paper are reshaped to the aspect ratio of the raw Viper [40] image for better visual presentation. More results on the label-to-image mapping comparison of UVIT (ours) and Improved ReCycleGAN are depicted in Figure 8 and the attached video *Compare.mp4*. For the *8_LRCompare.mp4*, we give a short description to guide the comparison. From left to right, there are outputs for four different input samples to compare:

- **1:** Please see the trajectory of the car and the surrounding road.
- **2:** Please see the boundary between two cars.
- **3:** Please see the translation of the road to check the complete translation and consistency across frames.
- **4:** Please see the consistency of the wall.

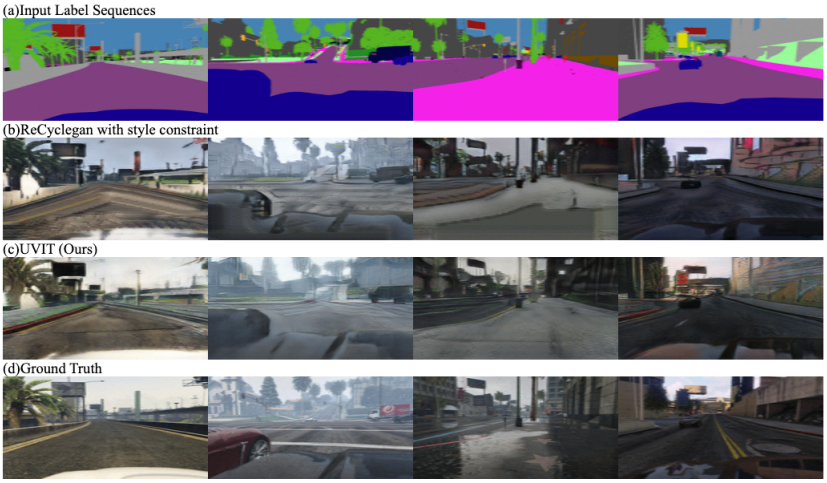

Figure 8. **Video screen cut of the label-to-image qualitative comparison.** First row: semantic label inputs; Second row: improved ReCycleGAN outputs; Third row: UVIT outputs. Fourth row: ground truth. A full video file can be found in *8_LRCompare.mp4*.

Figure 9. **Ablation study: when no sub-domain label is used during training and testing.** First video is the input semantic label sequence, the rest videos are the translated scene videos with style codes randomly sampled from prior distribution. There are 220 frames for each video. The corresponding video is attached as $9\_No\_subdomain.mp4$

4. Additional Results for Ablation Study

Here we provide the supplementary results for the ablation part. First, we provide complete quantitative experimental results that demonstrate the proposed video interpolation loss for a better V2V translation on both the image-to-label and the label-to-image tasks. Second, we study how the number of frames influence the semantic preserving performance. Third, we give the qualitative results of multimodal consistent videos when UVIT is trained and tested without the sub-domain label.

To feed more frames within a single GPU and compare with the main experiment in the main paper, we conduct the first and second ablation experiments with a resolution of 128×128.

To show more details within the existing GPU resource constraint, we conduct the third ablation experiment with a resolution of 256×256. The model is trained with 4 frames per batch. During the test time, we divide a longer sequence into sub-sequences of 4 frames with overlap.

4.1. UVIT without the subdomain label

To check how our UVIT performs in terms of style consistency without the subdomain label information, we run this ablation experiment. The results are attached in Figure 9 and $9\_No\_subdomain.mp4$. By randomly sampling the style code from prior distribution, we can get multimodal consistent video results in a stochastic way.

References

[1] Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented cyclegan: Learning many-to-many mappings from unpaired data. In: ICML. pp. 195–204 (2018)

[2] Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: Unsupervised video retargeting. In: ECCV. pp. 119–135 (2018)

[3] Bashkirova, D., Usman, B., Saenko, K.: Unsupervised video-to-video translation. arXiv preprint arXiv:1806.03698 (2018)

[4] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)

[5] 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. pp. 3213–3223 (2016)

[6] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)

[7] Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard gan. In: ICLR (2019)

[8] Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS. pp. 700–708 (2017)

[9] Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: ICCV. pp. 2213–2222 (2017)

[10] Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c.: Deep learning for precipitation nowcasting: A benchmark and a new model. In: NIPS. pp. 5617–5627 (2017)

[11] Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in Neural Information Processing Systems. pp. 1144–1156 (2018)

[12] Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)