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

Abstract

Existing unsupervised video-to-video translation methods fail to produce translated videos which are frame-wise realistic, semantic information preserving and video-level consistent. In this work, we propose UVIT, a novel unsupervised video-to-video translation model. Our model decomposes the style and the content, uses the specialized encoder-decoder structure and propagates the inter-frame information through bidirectional recurrent neural network (RNN) units. The style-content decomposition mechanism enables us to achieve style consistent video translation results as well as provides us with a good interface for modality flexible translation. In addition, by changing the input frames and style codes incorporated in our translation, we propose a video interpolation loss, which captures temporal information within the sequence to train our building blocks in a self-supervised manner. Our model can produce photo-realistic, spatio-temporal consistent translated videos in a multimodal way. Subjective and objective experimental results validate the superiority of our model over existing methods.

1. Introduction

Recent image-to-image translation (I2I) methods have achieved astonishing results by employing Generative Adversarial Networks (GANs) [18]. While there is an explosion of papers on I2I, its video counterpart is much less explored. Nevertheless, the ability to synthesize dynamic visual representations is important to a wide range of tasks (video colorization [51], medical imaging [36], model-based reinforcement learning [5, 21], computer graphics rendering [27], etc.).

Compared with the I2I task, the video-to-video translation (V2V) is more challenging. Besides the frame-wise realistic and semantic preserving requirements, which are also required in the I2I task, V2V methods additionally need to consider the temporal consistency for generating sequence-wise realistic videos. A straightforward idea towards this goal may be directly extending existing unsupervised I2I approaches with a video-based generator. However, video-based generators from other video tasks (*e.g.* video prediction [32, 43], unconditional video synthesis [37, 42]) are not born with the ability to produce frame-wise realistic and temporally consistent translated videos. Without elaborately designed training criterion, the generator will produce discrepant contents even with temporally continuous video inputs. To endow the generator with the ability of maintaining temporal consistency, the 3DCycleGAN method [7] introduce an 3D temporal discriminator to train a 3D sptio-temporal translator. However, as conducting adversarial training on high-dimensional video data is highly unstable, the 3D sptio-temporal translator often fails to obtain a good trade-off between image-level visual quality and video-level temporal consistency. To alleviate the training difficulty, the RecycleGAN [6] approach trains a separate temporal predictor to introduce explicit temporal constraint on the image-level translator. However, as the image-level translator alone could not incorporate multi-frame information during inference, inaccurate temporal prediction results often leads to flickering artifacts in the translation results.

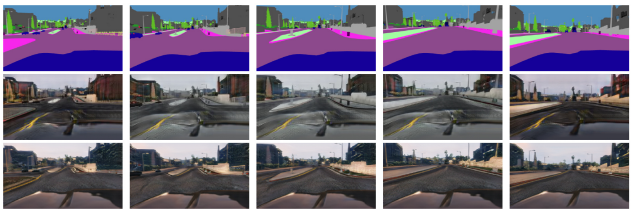

Figure 1. First row: label inputs; Second row: ReCycleGAN[6] outputs; Third row: UVIT (ours) outputs. To overcome the style shift (*e.g.* sunset frame gradually changes to rain frame), we utilize style-conditioned translation. To reduce artifacts across frames, our translator incorporate multi-frame information. We use systematic sampling to get the results from a 64-frame sequence.

In this paper we propose Unsupervised Multimodal Video-to-Video Translation via Self-Supervised Learning (UVIT), a novel framework for video-to-video cross-domain mapping. To this end, a temporally consistent video

*Equal contributions

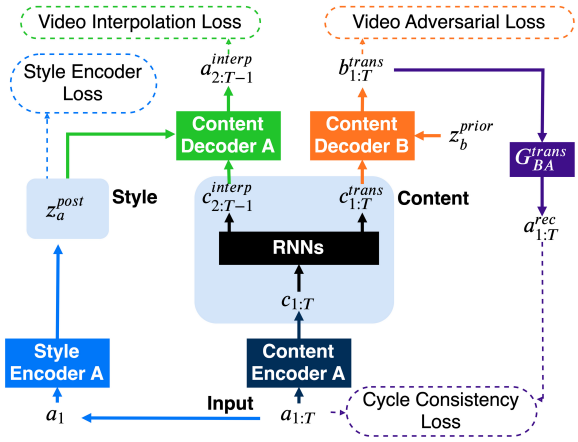

Figure 2. Overview of our proposed UVIT model: given an input video sequence, we first decompose it to the content by a Content Encoder and the style by a Style Encoder. Then the content is processed by special RNN units, namely TrajGRUs [43] to get the content used for translation and interpolation in a recurrent manner. Finally, the translation content and the interpolation content are decoded to the translated video and the interpolated video together with the style latent variable. We also show the video adversarial loss (orange), the cycle consistency loss (violet), the video interpolation loss (green) and the style encoder loss (blue)

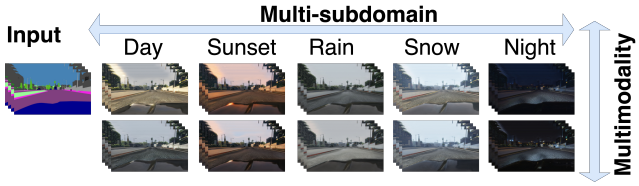

Figure 3. Our proposed UVIT model can produce photo-realistic, spatio-temporal consistent translated videos in a multimodal way for multiple subdomains

sequence translation should simultaneously guarantee: *(1)* Style consistency, and *(2)* Content consistency, see Figure 1 for a visual example. Style consistency requires the whole video sequence to have the same style, thus ensuring the video frames to be overall realistic. Meanwhile, content consistency refers to the appearance continuity of contents in adjacent video frames, which ensures the video sequence to be dynamically vivid.

In UVIT, by assuming that all domains share the same underlying structure, namely content space, we exploit the style-conditioned translation. To simultaneously impose style and content consistency, we adopt an Encoder-RNN-Decoder architecture as the video translator, see Figure 2 for an illustration of the proposed framework. There are two key ingredients in our framework:

Conditional video translation: By applying the same style code to decode the content feature for a specific translated video, the translated video is style consistent. Besides, by changing the style code across videos, we achieve subdomain[1] and modality flexible video translation, see Figure 3 for an illustration of subdomains (columns) and modalities (rows).

Consistent video translation: Besides introducing style code to ensure style consistency, another more significant contribution of our paper lies in our self-supervised RNN translator. To mitigate the difficulties in unsupervised training, existing unsupervised V2V approaches utilize simple generators such as 3D Convolutional Neural Networks (CNNs) or 2D CNN + temporal predictor. However, these translators have limited capacity in capturing complex relationship between multiple video frames. In this paper, inspired by its success in other video processing tasks [43], we adopt a RNN-based translator to exploit temporal information from more frames. Training RNN to extract meaningful latent representations is a non-trivial task even with supervised losses [20, 10], in our unsupervised framework, we propose to make full use of video data and introduce a video interpolation loss to train our RNN building blocks in a self-supervised manner. Specifically, we utilize the latent representations in our RNN to perform the video translation and video interpolation tasks simultaneously. The pixel-level supervised interpolation loss endows our RNN with the ability of maintaining temporal continuity and greatly stabilizes the challenging unpaired video adversarial learning. It is worth noting that existing video self-supervised learning auxiliary tasks (*e.g.* video ordering [34, 16, 30, 29] and statistics prediction [45]) are mainly designed for high-level video understanding (*e.g.* video classification, action recognition), while our interpolation loss introduces pixel-wise supervision for the challenging unsupervised video translation task.

The main contributions of our paper are summarized as follows:

1. We propose an unsupervised video to video (V2V) translation framework, which decomposes the temporal consistency into style and content consistencies for stable and coherent video translation.
2. We propose an Encoder-RNN-Decoder video generator, the style-conditioned decoder ensures style consistency as well as facilitates multimodal video translation.
3. We propose an innovative video interpolation loss which introduces pixel-wise supervision to train our RNN generator in a self-supervised manner.

2. Related Work

Image-to-Image Translation. Most of the GAN-based I2I methods mainly focus on the case where paired data exists [25, 53, 47]. However, with the cycle-consistency loss

[1]Hereafter we call it *subdomain* and not domain because a subdomain must belong to a subset of a domain (for instance, subdomains of day, night, snow, etc. belong to the scene video domain)

introduced in CycleGAN [52], promising performance has been achieved also for the unsupervised I2I [24, 2, 31, 35, 41, 17, 13, 49, 12, 48, 1]. The conditional distribution of the translated pictures on the input pictures is quite likely to be multimodal (*e.g*. from a semantic label to different images in a fixed weather condition). However, traditional I2I problem often lacks this characteristic and produces an unimodal outcome. Zhu *et al*. [53] proposed Bicycle-GAN that can output diverse translations in a supervised manner. There are also some extensions [24, 2, 28] of CycleGAN to decompose the style and content so that the output can be multimodal in the unsupervised scenario. Our work goes in this direction, and under the assumption that close frames within the same domain share the same style, we adopt the style control strategy in the image domain proposed by Almahairi *et al*. [2] to the video domain.

Video-to-Video Translation In the seminal work, Wang *et al*. [46] (vid2vid) combined the optical flow and video-specific constraints and proposed a general solution for V2V in a supervised way, which achieves long-term high-resolution video sequences. However, vid2vid relies heavily on labeled data which makes it difficult to scale in unsupervised real-world scenarios.

To deal with the more challenging unsupervised V2V task, recent methods [7, 6, 11] extend the I2I CycleGAN approach by employing spatio-temporal loss to introduce extra temporal constraint. Bashkirova *et al*. [7] proposed a 3DCycleGAN method which adopts 3D convolutions in the generator and discriminator of the CycleGAN framework to capture temporal information. However, since the small 3D convolution operator (with a small temporal dimension 3) only captures dependency between adjacent frames. 3DCycleGAN can't exploit temporal information for generating longer style consistent video sequences. Furthermore, conduct adversarial training on high-dimensional video tensor is highly unstable, as a result, 3DCycleGAN tends to sacrifice the image-level quality and generates blurry and gray translations.

Additionally, Bansal *et al*. [6] designed a recycle loss (ReCycleGAN) for jointly modeling the spatio-temporal relationship between video frames. They trained a temporal predictor to predict the next frame based on two past frames, and plugged the temporal predictor in the cycle-loss to impose the spatio-temporal constraint on the traditional image-level translator. Although ReCycleGAN succeeds in V2V translation scenarios such as face-to-face or flower-to-flower, similar to CycleGAN, it lacks domain generalization as the translation fails to be consistent in domains with a large gap with respect to the input. We argue that there are two major reasons that affect ReCycleGAN performance in complex scenarios. First, its image-level translator processes input frames independently, has limited capacity in exploiting temporal information, being not content consistent enough. Second, ReCycleGAN temporal predictor only imposes the temporal constraint between a few adjacent frames, the generated video content still might shift abnormally: a sunny scene could change to a snowy scene in the following frames. Note that Chen *et al*. [11] incorporate optical flow to add motion cycle consistency and motion translation constraints. However, their Motion-guided CycleGAN still suffers from the same two limitations as in ReCycleGAN.

In summary, previous methods fail to produce style consistent and multimodal video sequences. Besides, they lack the ability to achieve translation which is both content consistent enough and frame-wise realistic. In this paper, we propose UVIT, a novel method for Unsupervised Multimodal Video-to-Video Translation via Self-Supervised Learning, which produces high-quality semantic preserving frames with consistency within the video sequence. Besides, to the best of our knowledge, our method is the first method that jointly addresses multiple-subdomains and multimodality in V2V cross-domain translations.

3. UVIT for Unsupervised V2V Translation

3.1. Problem setting

Let A be the video domain A, $a_{1:T} = \{a_1, a_2, ..., a_T\}$ be a sequence of video frames in A, let B be the video domain B, $b_{1:T} = \{b_1, b_2, ..., b_T\}$ be a sequence of video frames in B. For example, they can be sequences of semantic segmentation labels or scene images. Our general goal of unsupervised video-to-video translation is to train a generator to convert videos between domain A and domain B with many-to-many mappings. Either domain A or domain B can have multiple subdomains (sunny, snow, rain for the case of weather conditions). More concretely, to generate the style consistent video sequence, we assume each video frame has a shared style latent variable z. Let $z_a \in Z_A$ and $z_b \in Z_B$ be the style latent variables in domain A and B, respectively.

We aim to achieve two conditional video translation mappings: $\mathbb{G}_{AB}^{trans} : A \times Z_B \mapsto B^{trans}$ and $\mathbb{G}_{BA}^{trans} : B \times Z_A \mapsto A^{trans}$. As we propose to use the video interpolation loss to train the translator components in a self-supervised manner, we also define the video interpolation mappings: $\mathbb{G}_{A}^{interp} : A \times Z_A \mapsto A^{interp}$ and $\mathbb{G}_{B}^{interp} : B \times Z_B \mapsto B^{interp}$. Interpolation and translation mappings use exactly the same building blocks.

3.2. Translation and Interpolation pipeline

In this work, inspired by UNIT [31], we assume a shared content space such that corresponding frames in two domains are mapped to the same latent content representation. We show the translation and interpolation processes in Figure 4. To achieve the goal of unsupervised video-to-video

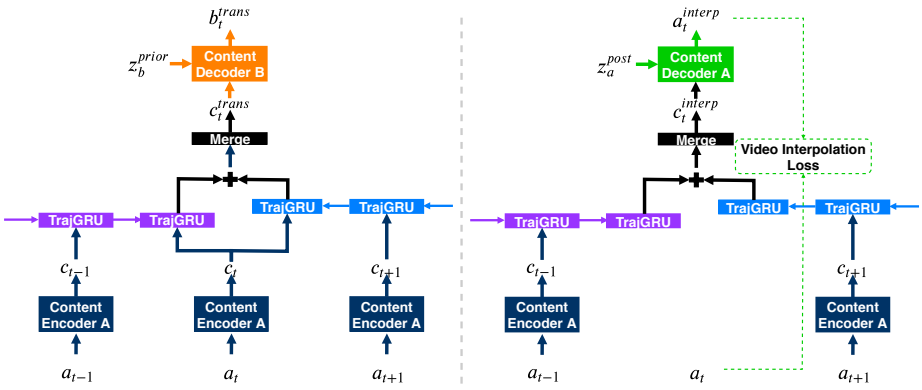

Figure 4. Video translation (left) and video interpolation (right): two processes share modules organically. The input latent content is processed by the Merge Module to merge information from TrajGRUs in both the forward and the backward direction. The translation content (c_t^{trans}) is obtained by updating interpolation content (c_t^{interp}) with the content (c_t) from the current frame (a_t)

translation, we propose an Encoder-RNN-Decoder translator which contains the following components:

- Two content encoders ($\mathbb{CE}_A$ and $\mathbb{CE}_B$): extract the frame-wise content information from each domain to the common spatial content space.
- Two style encoders ($\mathbb{SE}_A$ and $\mathbb{SE}_B$): encode video frames to the respective style domains.
- Two Trajectory Gated Recurrent Units (TrajGRUs) [43] to form a Bi-TrajGRU ($\mathbb{T}$): propagate the inter-frame content information bidirectionally. TrajGRU [43] is one variant of Convolutional RNN (Recurrent Neural Network) [50], which can actively learn the location-variant structure in the video data. More details in supplementary material.
- One merge module ($\mathbb{M}$): adaptively combine the inter-frame content from two directions.
- Two conditional content decoders ($\mathbb{CD}_A$ and $\mathbb{CD}_B$): take the spatio-temporal content information and the style code to generate the output frame. If needed, it also takes the conditional subdomain information as a one hot vector encoding.

Video translation: Given an input frame sequence $(..., a_{t-1}, a_t, a_{t+1}, ...)$, we extract the posterior style (z_a^{post}) from the first frame (a_1) with a style encoder ($\mathbb{SE}_A$). Additionally, we extract each content representation $(..., c_{t-1}, c_t, c_{t+1}, ...)$ with the content encoder ($\mathbb{CE}_A$).

Translation is conducted in a recurrent way. To get the translation result b_t^{trans} for time t , we process the independent content representation: (1) propagate content for the surrounding frames $(..., c_{t-1}, c_{t+1}, ...)$ through *Bi-TrajGRU* ($\mathbb{T}$) to obtain the inter-frame content information. (2) update this information with the current frame content (c_t) (see Figure 4 left, Merge Module $\mathbb{M}$) to get the spatio-temporal content (c_t^{trans}) for translation. At last, using the same style-conditioned strategy as Augment CycleGAN [2, 15, 38], the content decoder ($\mathbb{CD}_B$) takes the prior style information (z_b^{prior}) drawn from the prior distribution as the condition and utilizes c_t^{trans} to generate the translation result ($\mathbb{CD}_B(c_t^{trans}, z_b^{prior}) = b_t^{trans}$). This process is repeated until we get the whole translated sequence (...,b_{t-1}^{trans}, b_t^{trans},b_{t+1}^{trans},...).

The style code is induced as the condition of (AdaIN-based [23]) content decoder. If a domain (*e.g.* scene images) is presorted, we have prior information on which subset (rain, night, etc.) a video belongs to and can take such prior information as a subdomain (subset) label to achieve deterministic control for style. Within each subset, there are still different modalities (*e.g.* overcast day, sunny day in day subdomain), yet we don't have prior access to it. This modality information is hence learned by style encoder. Subdomain label (taken as one-hot vector if available) and modality information together constitute 21-dimensional style code. Style consistency is ensured by sharing style code among a specific video sequence. Multimodal translation is realized by inducing different style codes across videos. When subdomain information is unavailable, simply using style encoder to learn subdomain styles as modalities, we can still generate multimodal style consistent results stochastically.

Video interpolation: In video translation process in Figure 4, when translating a specific frame (a_t), the translation content (c_t^{trans}) is integrated by the current frame content (c_t) and inter-frame information (c_t^{interp}) from the surrounding frames $(..., a_{t-1}, a_{t+1}, ...)$. Inter-frame content information helps to build up the dependency between each frame and its surrounding frames, ensuring content consistency across frames. However, if c_t^{interp} and c_t are not aligned well, image-level quality can be affected. The translated frame (b_t^{trans}) will incline to over smooth image-level details and sacrifice high-level semantic correspondence with a_t. Tailoring inter-frame information is thus of pivotal importance. Thanks to the flexible Encoder-

Decoder structure, our decoder can generate the interpolated frame (a_t^{interp}) from c_t^{interp}. Video interpolation loss is proposed to compute the L1 distance between interpolated frame(a_t^{interp}) and the current frame(a_t), which directly adds supervision to the inter-frame content(c_t^{interp}). Therefore, the translation task directly benefits from the interpolation task to obtain the ability of maintaining temporal consistency.

Meanwhile, such self-supervised training would be beneficial to make the network more stable in the challenging unpaired video adversarial learning, as shown in [44]. GANs are powerful methods to learn a data probability distribution with no supervision, yet training GANs is well known for being delicate, unstable [3, 33, 4] and easy to suffer from mode collapse[6]. Besides cycle loss acting as spatial constraint, we introduce the video interpolation loss as a temporal constraint for GAN training in a self-supervised way. It has been validated that bringing self-supervision is beneficial for cross-domain unsupervised tasks (*e.g.* natural image synthesis) [44, 9, 39]. What's more, our framework aims to learn latent representation for style and content, while it has been empirically observed [20, 10] that it is non-trivial to extract meaningful latent representations from observed sequence when coupled with a strong autoregressive decoder (*e.g.* RNN). Goyal *et al*. [19] found that auxiliary cost could ease training of the latent variables in RNN-based generative latent variable models. Therefore, the video interpolation task provides the latent variables with a auxiliary objective that enhances the performance of the overall model.

Note that the proposed temporal loss highly differs from the previous ReCycleGAN loss [6] as: *(1)* we use a RNN-based architecture that captures temporal information better in a high-level feature space, *(2)* interpolation is conducted within the translator building blocks rather than using different modules, training all the translator building blocks with direct self-supervision, *(3)* the translator directly utilizes tailored inter-frame information for better semantic preserving translations.

3.3. Loss functions

We use the Relativistic GAN (RGAN) [26] and the least square [33] version for the adversarial loss. RGAN estimates the probability that the given real data is more realistic than a randomly sampled fake data. We use image-level discriminators (D_x^{img}) and video-level (D_x^{vid}) discriminators to ensure that output frames resemble a real video clip in both video-level and image-level. Moreover, we also add style discriminators (D_{Z_x}) to adopt an adversarial approach for training style encoders.

Video adversarial loss. The translated video frames aim to be realistic compared to the real samples in the target domain for both an image-level and a video-level basis.

$$L_B^{adv} = \frac{1}{T}\sum\nolimits_{i=1}^{i=T}[D_B^{img}(b_i^{trans}) - D_B^{img}(b_i) - 1]^2 + [D_B^{vid}(b_{1:T}^{trans}) - D_B^{vid}(b_{1:T}) - 1]^2, \tag{1}$$

where, $b_{1:T}^{trans}$ are the translated frames from time 1 to T. D_B^{img} and D_B^{vid} are the image-level and video-level discriminators for domain B. Adversarial loss for domain A (L_A^{adv}) is defined similarly.

Video interpolation loss. The interpolated video frames should be close to the target frames (pixel-wise loss), and be realistic compared to other real frames within the domain (adversarial loss).

$$L_A^{interp} = \frac{1}{(T-2)}(\lambda_{interp} \parallel a_{2:T-1} - a_{2:T-1}^{interp} \parallel_1 + \sum\nolimits_{i=2}^{i=T-1}[D_A^{img}(a_i^{interp}) - D_A^{img}(a_i) - 1]^2). \tag{2}$$

Since we are using bidirectional TrajGRUs, we use frames from time 2 to $T-1$ to compute the video interpolation loss. $a_{2:t-1}^{interp}$ are the interpolated frames. The first part of the loss is the supervised pixel-wise L_1 loss, and the later part is the GAN loss computed on the image-level discriminator D_A^{img}. λ_{interp} is used to control the weight between two loss elements.

Cycle consistency loss. In order to ensure semantic consistency in an unpaired setting, we use a cycle-consistency loss:

$$L_A^{cycle} = \frac{\lambda_{cycle}}{T} \parallel a_{1:T} - a_{1:T}^{rec} \parallel_1, \tag{3}$$

where $a_{1:T}^{rec}$ are the reconstructed frames of domain A from time 1 to T, *i.e.* $a_{1:T}^{rec} = \mathbb{G}_{BA}^{trans}(b_{1:T}^{trans}, z_a^{post})$. Where z_a^{post} is the posterior style variable produced by using the style encoder to encode a_1. λ_{cycle} is the cycle consistency loss weight.

Style encoder loss. To train the style encoder, the style reconstruction loss and style adversarial loss are defined in a similar way as Augmented CycleGAN [2]:

$$L_{Z_A}^{style} = \lambda_{rec} \parallel z_a^{rec} - z_a^{prior} \parallel_1 + [D_{Z_A}(z_a^{post}) - D_{Z_A}(z_a^{prior}) - 1]^2. \tag{4}$$

Here, z_a^{prior} is the prior style latent variable of domain A drawn from the prior distribution. z_a^{rec} is the reconstructed style latent variable of domain A by using the style encoder to encode a_1^{trans}. λ_{rec} is the style reconstruction loss weight.

Therefore, the objective for the generator is:

$$L_G^{total} = L_A^{adv} + L_B^{adv} + L_A^{interp} + L_B^{interp} + L_A^{cycle} + L_B^{cycle} + L_{Z_A}^{style} + L_{Z_B}^{style} \tag{5}$$

Detailed λ values and loss functions for discriminators are attached in the supplementary material. Detailed training algorithm for RGANs can be found in [26].

4. Experiments

We validate our method using two common yet challenging datasets: Viper [40], and Cityscapes [14] datasets. We conduct image-to-label, label-to-image and cross subdomain translation on Viper [40], and also translate videos between Viper and Cityscapes. To feed more frames within a single GPU, we use image with 128×128 and 10 frames per batch for the main experiments. During inference, we use video sequences of 30 frames. These 30 frames are divided into 4 smaller sub-sequences of 10 frames with overlap. They all share the same style code to be style consistent. Note that our model can be easily extended to process longer style-consistent sequences by sharing the same style code for the sub-sequences. The example of longer style consistent video is provided in supplementary material, where detailed description of the dataset and implementation are also attached.

4.1. Ablation Study

In order to demonstrate the contribution of our method, we first conduct ablation study experiments. We provide quantitative and qualitative experimental results that demonstrate the proposed video interpolation loss for a better V2V translation. Besides, we provide results of how number of frames influences the semantic preserving performance. We also provide multimodal consistent results of model trained without using subdomain label in the supplementary material.

Video interpolation loss. We provide ablation experiments to show the effectiveness of the proposed video interpolation loss. We conduct experiments on both the image-to-label and the label-to-image tasks. We denote UVIT trained without video interpolation loss as "UVIT wo/vi".

We follow the experimental setting of ReCycleGAN [6] and use semantic segmentation metrics to quantitatively evaluate the image-to-label results. The Mean Intersection over Union (mIoU), Average Class Accuracy (AC) and Pixel Accuracy (PA) scores for ablation experiments are reported in Table 1. Our model with video interpolation loss achieves the best performance across subdomains, which confirms that the video interpolation helps to preserve the semantic information between the translated frame and the corresponding input frame.

The Fréchet Inception Distance (FID) [22] was originally developed for image generation evaluation. Wang *et al*. propose a variant – FID for video to evaluate video, which measures both visual quality and temporal consistency. Specifically, FID for video uses the pre-trained video recognition CNN as a feature extractor. For the label-to-image task, we use the FID for video [46] to evaluate the feature distribution distance between translated videos and ground-truth videos. Similar to vid2vid [46], we use the pre-trained network (I3D [8]) to extract spatio-temporal features from video sequences. We extract the semantic labels from the respective sub-domains to generate videos and evaluate the FID score on all the subdomains of the Viper dataset. Table 2 shows the FID score for UVIT and the corresponding ablation experiment. On both the image-to-label and label-to-image tasks, the proposed video interpolation loss plays a crucial role for UVIT to achieve good translation results.

Table 1. **Image-to-Label (Semantic segmentation) quantitative evaluation.** We validate UVIT without video interpolation loss (wo/vi) under Mean Intersection over Union (mIoU), Average Class Accuracy (AC) and Pixel Accuracy (PA) scores

Criterion	Model	Day	Sunset	Rain	Snow	Night
mIoU ↑	UVIT wo/vi	10.14	10.70	11.06	10.30	9.06
	UVIT	**13.71**	**13.89**	**14.34**	**13.23**	**10.10**
AC ↑	UVIT wo/vi	15.07	15.78	15.46	15.01	13.06
	UVIT	**18.74**	**19.13**	**18.98**	**17.81**	**13.99**
PA ↑	UVIT wo/vi	56.33	57.16	58.76	55.45	55.19
	UVIT	**68.06**	**66.35**	**67.21**	**65.49**	**58.97**

Table 2. **Label-to-image quantitative evaluation.** We validate our system without video interpolation loss (wo/vi) under the Fréchet Inception Distance (FID) score

Criterion	Model	Day	Sunset	Rain	Snow	Night
FID ↓	UVIT wo/vi	26.95	23.04	30.48	34.62	47.50
	UVIT	**17.32**	**16.79**	**19.52**	**18.91**	**19.93**

Table 3. **Quantitative results of UVIT with different number of frames per batch in training on the image-to-label (Semantic segmentation) task.** With the increase of input frames number in the sub-sequence, our RNN-based translator can utilize the temporal information better, resulting in better semantic preserving

Criterion	Frame number	Day	Sunset	Rain	Snow	Night	All
mIoU↑	4	11.84	11.91	12.35	11.37	8.49	11.19
	6	12.29	12.66	13.03	11.77	9.79	11.94
	8	13.05	13.21	14.23	13.07	**11.00**	12.87
	10	**13.71**	**13.89**	**14.34**	**13.23**	10.10	**13.07**
AC↑	4	16.78	16.75	16.57	16.32	12.21	15.7
	6	17.50	17.46	17.66	16.73	14.23	16.62
	8	18.42	18.28	**19.19**	17.80	**15.18**	**17.68**
	10	**18.74**	**19.13**	18.98	**17.81**	13.99	17.59
PA↑	4	62.84	60.34	61.97	58.77	51.68	59.04
	6	62.85	61.21	62.21	59.77	56.84	60.51
	8	65.56	64.11	66.26	64.18	**62.02**	64.25
	10	**68.06**	**66.35**	**67.21**	**65.49**	58.97	**65.20**

Different number of input frame. Our RNN-based translator incorporates temporal information from multiple frames. We also investigate the influence of frame number on the performance of our model. As shown in Table 3 UVIT can achieve better semantic preserving with more frames feeding during training as the RNNs are better trained to leverage the temporal information. Specifically, for the image-to-lable translation, with the increase of the number from 4 to 10, the overall mIoU increase from 11.19 to 13.07.

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

Criterion	Model	Day	Sunset	Rain	Snow	Night	All
mIoU↑	Cycle-GAN	3.39	3.82	3.02	3.05	7.76	4.10
	ReCycleGAN (Reproduced)[1]	10.31	11.18	11.26	9.81	7.74	10.11
	UVIT (Ours)	**13.71**	**13.89**	**14.34**	**13.23**	**10.10**	**13.07**
AC↑	Cycle-GAN	7.83	8.56	7.91	7.53	11.12	8.55
	ReCycleGAN (Reproduced)[1]	15.78	15.80	15.95	15.56	11.46	14.84
	UVIT (Ours)	**18.74**	**19.13**	**18.98**	**17.81**	**13.99**	**17.59**
PA↑	Cycle-GAN	15.46	16.34	12.83	13.20	49.03	19.59
	ReCycleGAN (Reproduced)[1]	54.68	55.91	57.72	50.84	49.10	53.65
	UVIT	**68.06**	**66.35**	**67.21**	**65.49**	**58.97**	**65.20**

Table 5. **Quantitative comparison between UVIT and baseline approaches on label-to-image task.** Better FID indicates that our translation has better visual quality and temporal consistency

Criterion	Model	Day	Sunset	Rain	Snow	Night
FID↓	ReCycleGAN [6]	23.60	24.45	28.54	31.58	35.74
	ReCyclegan with style constraint	20.39	21.32	25.67	21.44	21.45
	UVIT (ours)	**17.32**	**16.79**	**19.52**	**18.91**	**19.93**

4.2. Comparison of UVIT with State-of-the-Art Methods

Image-to-label mapping. To further ensure reproducibility, we use the same setting as our ablation study to compare UVIT with ReCycleGAN [6] in the image-to-label mapping task. We report the mIoU, AC and PA metrics by the proposed approach and competing methods in Table 4. The results clearly validate the advantage of our method over the competing approaches in terms of preserving semantic information. Our model can effectively leverage the inter-frame information from more frames in a direct way, which utilizes the temporal information better than the indirect way in ReCycleGAN [6].

Label-to-image mapping. In this setting, we compare the quality of the translated video sequence by different methods. We extract the semantic labels from the respective sub-domains to generate videos and evaluate the FID score for videos on all the subdomains of the Viper dataset in the same setting as our ablation experiments. As shown in Figure 1, the original ReCycleGAN output video sequences can not ensure style consistency. We also report the results achieved by our improved version of the ReCycleGAN for reference. Concretely, we develop a conditional version which formally controls the style of generated video sequences in a similar way as our UVIT model, and denote the conditional version as ReCyclegan with style constraint. The FID results by different methods are shown in Table 5. The proposed UVIT achieves better FID on all the 5 sub-domains, which validates the effectiveness of our model. Combining Table 2 and Table 5, there is another observation – the UVIT w/o vi-loss could not dominate the Improved ReCycleGAN in terms of FID. This shows that the video interpolation loss is crucial for the superiority of our spatio-temporal translator.

To thoroughly evaluate the visual quality of the video translation results, we conduct subjective evaluation on the Amazon Mechanical Turk (AMT) platform. Detailed information of conducting this subjective test is provided in the supplementary material. We compare the proposed UVIT with 3DCycleGAN and ReCycleGAN. The video-level and image-level human preference scores (HPS) are reported in Table 6. For reference, we also compare the video-level quality between UVIT and the supervised vid2vid model [46]. Meanwhile, image-level quality comparison between UVIT and CycleGAN (the image translation baseline using 2D CNN) is also included. Table 6 clearly demonstrates the effectiveness of our proposed UVIT model. In video-level comparison, our unsupervised UVIT model outperforms the competing unsupervised ReCycleGAN and 3DCycleGAN by a large margin, and achieves comparable results with the supervised benchmark. In image-level comparison, UVIT achieves better HPS than both V2V competing approaches and image-to-image baseline. For a better comparison, we include several qualitative examples of generated videos in supplementary material.

4.3. More experimental results

High resolution results. To get a higher resolution and show more details within the existing GPU constraint, we

[1]Output would be in a resolution of 256×256, we then downscale it to 128×128 to compute statistics.

Table 6. **Label-to-image Human Preference Score.** UVIT outperforms all the competing unsupervised methods. Note that we achieve comparable performance with vid2vid although it is supervised

Human Preference Score	Video level	Image level
UVIT (ours) / ReCyclegan with style constraint	**0.67** / 0.33	**0.66** / 0.34
UVIT (ours) / 3DCycleGAN [7]	**0.75** / 0.25	**0.70** / 0.30
UVIT (ours) / vid2vid [46]	0.49 / **0.51**	–
UVIT (ours) / CycleGAN [52]	–	**0.61** / 0.39

Figure 5. **Viper rain-and-snow.** From left to right: input rain video, translated snow video, input snow video, translated rain video. Corresponding videos are provided in supplementary material

Figure 6. **Viper sunset-and-day** From left to right: input sunset video, translated day video, input day video, translated sunset video.

Figure 7. **Cityscapes-to-Viper.** From left to right: Cityscapes video, translated Viper night video, translated Viper snow video, translated Viper sunset video

also train our model using images of 256×256 and 4 frames per batch, then test with longer sequence, which is divided to subsequences of 4 frames with overlap. A visual example is shown in Figure 1. Note that all visual examples in this paper are reshaped to the aspect ratio of the raw Viper [40] image for better visual presentation. Quantitatively results, visual examples and videos are provided in supplementary material.

Translation on other datasets. Besides translating video sequences between image and semantic label domains, we also train models to translate video sequences between different scene image subdomains and different video datasets.

In Figure 5, we provide visual examples of translation between Rain and Snow scenes in the Viper dataset. In Figure 6, we provide visual examples of translation between Sunset and Day scenes. Visual examples of translation between Viper and Cityscapes [14] datasets is organized in figure 7. They show the ability of our approach to learn the association between synthetic videos and real-world videos. More examples and the corresponding videos are attached in supplementary material.

5. Conclusion

In this paper, we have proposed UVIT, a novel method for unsupervised video-to-video translation. A specialized Encoder-RNN-Decoder spatio-temporal translator has been proposed to decompose style and content in the video for temporally consistent and modality flexible video-to-video translation. In addition, we have designed a video interpolation loss within the translator which utilizes highly structured video data to train our translators in a self-supervised manner. This enables the effective application of RNN-based network in the challenging V2V task. Extensive experiments have been conducted to show the effectiveness of the proposed UVIT model. Without using any paired training data, the proposed UVIT model is capable of producing excellent multimodal video translation results, which are image-level realistic, semantic information preserving and video-level consistent.

Acknowledgments. This work was partly supported by the ETH Zürich Fund (OK), and by Huawei, Amazon AWS and Nvidia grants.

References

[1] Alharbi, Y., Smith, N., Wonka, P.: Latent filter scaling for multimodal unsupervised image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1458–1466 (2019)

[2] 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)

[3] Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: ICLR (2017)

[4] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. In: ICML. pp. 214–223 (2017)

[5] Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866 (2017)

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

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

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

[9] Chen, T., Zhai, X., Ritter, M., Lucic, M., Houlsby, N.: Self-supervised gans via auxiliary rotation loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12154–12163 (2019)

[10] Chen, X., Kingma, D.P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., Abbeel, P.: Variational lossy autoencoder. arXiv preprint arXiv:1611.02731 (2016)

[11] Chen, Y., Pan, Y., Yao, T., Tian, X., Mei, T.: Mocycle-gan: Unpaired video-to-video translation. In: ACM International Conference on Multimedia. pp. 647–655. ACM (2019)

[12] Cho, W., Choi, S., Park, D.K., Shin, I., Choo, J.: Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10639–10647 (2019)

[13] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)

[14] 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)

[15] Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: ICLR (2017)

[16] Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3636–3645 (2017)

[17] Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. In: CVPR. pp. 2477–2486 (2019)

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)

[19] Goyal, A.G.A.P., Sordoni, A., Côté, M.A., Ke, N.R., Bengio, Y.: Z-forcing: Training stochastic recurrent networks. In: Advances in neural information processing systems. pp. 6713–6723 (2017)

[20] Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A.A., Visin, F., Vazquez, D., Courville, A.: Pixelvae: A latent variable model for natural images. arXiv preprint arXiv:1611.05013 (2016)

[21] Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)

[22] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS. pp. 6626–6637 (2017)

[23] Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017)

[24] Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV. pp. 172–189 (2018)

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

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

[27] Kajiya, J.T.: The rendering equation. In: ACM SIGGRAPH computer graphics. vol. 20, pp. 143–150. ACM (1986)

[28] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)

[29] Kim, D., Cho, D., Kweon, I.S.: Self-supervised video representation learning with space-time cubic puzzles. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8545–8552 (2019)

[30] Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676 (2017)

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

[32] Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)

[33] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. pp. 2794–2802 (2017)

[34] Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)

[35] Mo, S., Cho, M., Shin, J.: Instagan: Instance-aware image-to-image translation. In: ICLR (2019)

[36] Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D.: Estimating ct image from mri data using 3d fully convolutional networks. In: Deep Learning and Data Labeling for Medical Applications, pp. 170–178. Springer (2016)

[37] Ohnishi, K., Yamamoto, S., Ushiku, Y., Harada, T.: Hierarchical video generation from orthogonal information: Optical flow and texture. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

[38] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (2018)

[39] Ren, Z., Jae Lee, Y.: Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 762–771 (2018)

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

[41] Romero, A., Arbeláez, P., Van Gool, L., Timofte, R.: Smit: Stochastic multi-label image-to-image translation. In: ICCV Workshops. pp. 0–0 (2019)

[42] Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2830–2839 (2017)

[43] 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)

[44] Sun, Y., et al.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)

[45] Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4006–4015 (2019)

[46] 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)

[47] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR. pp. 8798–8807 (2018)

[48] Wu, P.W., Lin, Y.J., Chang, C.H., Chang, E.Y., Liao, S.W.: Relgan: Multi-domain image-to-image translation via relative attributes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5914–5922 (2019)

[49] Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: Geometry-aware unsupervised image-to-image translation. In: Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition. pp. 8012–8021 (2019)

[50] 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)

[51] Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8052–8061 (2019)

[52] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)

[53] Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NIPS. pp. 465–476 (2017)