

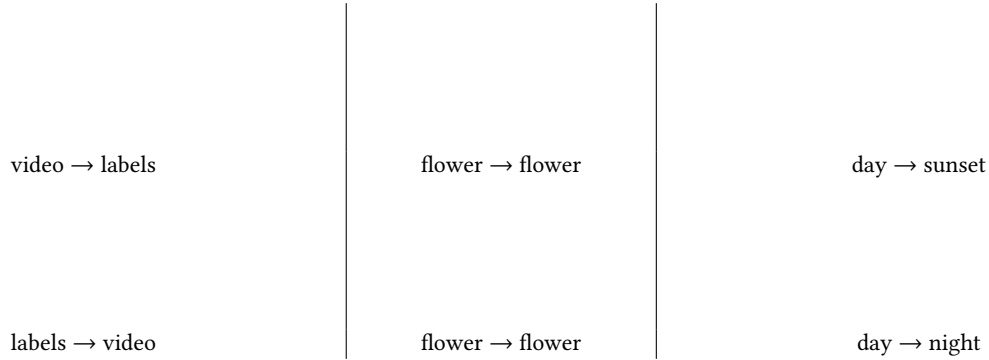
# Mocycle-GAN: Unpaired Video-to-Video Translation \*

Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian and Tao Mei

University of Science and Technology of China, Hefei, China

JD AI Research, Beijing, China

cheny01@mail.ustc.edu.cn;{panyw.ustc,tingyao.ustc}@gmail.com;xinmei@ustc.edu.cn;tmei@jd.com



**Figure 1:** Given any two unpaired video collections  $X$  and  $Y$ , our Mocycle-GAN learns to translate videos from source domain  $X$  to target domain  $Y$ . Left: Translation between Game scene videos and segmentation label maps. Center: Translation between time-lapse videos of variant flowers. Right: Translation between Game scene videos under different ambient conditions, e.g., rendering day-light video to the sunset/night environments. The *animated videos* are best viewed via Adobe Acrobat.

## ABSTRACT

Unsupervised image-to-image translation is the task of translating an image from one domain to another in the absence of any paired training examples and tends to be more applicable to practical applications. Nevertheless, the extension of such synthesis from image-to-image to video-to-video is not trivial especially when capturing spatio-temporal structures in videos. The difficulty originates from the aspect that not only the visual appearance in each frame but also motion between consecutive frames should be realistic and consistent across transformation. This motivates us to explore both appearance structure and temporal continuity in video synthesis. In this paper, we present a new Motion-guided Cycle GAN, dubbed as Mocycle-GAN, that novelly integrates motion estimation into unpaired video translator. Technically, Mocycle-GAN capitalizes on three types of constrains: adversarial constraint discriminating between synthetic and real frame, cycle consistency encouraging an inverse translation on both frame and motion, and motion translation validating the transfer of motion between consecutive frames. Extensive experiments are conducted on video-to-labels and labels-to-video translation, and superior results are reported when

comparing to state-of-the-art methods. More remarkably, we qualitatively demonstrate our Mocycle-GAN for both flower-to-flower and ambient condition transfer.

## CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → *Vision for robotics*; *Motion capture*.

## KEYWORDS

Video-to-Video Translation; GANs; Unsupervised Learning

### ACM Reference Format:

Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian and Tao Mei. 2019. Mocycle-GAN: Unpaired Video-to-Video Translation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3343031.3350937>

## 1 INTRODUCTION

The development of deep learning has led to a significant surge of research activities for multimedia content generation in multimedia and computer vision community. In between, image-to-image translation is one of the widely studied tasks and the recent advances in Generative Adversarial Networks (GANs) have successfully obtained remarkable improvements on image translation across domains. The achievements of image-to-image translation are on the assumption that a large amount of annotated and matching image pairs are accessible for model training. In practice, nevertheless, the manual labeling of such paired data is cost-expensive and even unrealistic. To address this issue, [15, 18, 35, 37] tackle image-to-image translation in an unsupervised manner, which only capitalizes on unpaired data (i.e., two sets of unlabeled images from two domains). In this paper, we go one step further and extend such synthesis

\*This work was performed at JD AI Research.

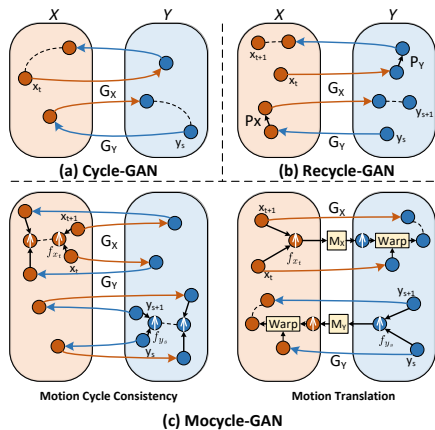
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '19, October 21–25, 2019, Nice, France

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6889-6/19/10...\$15.00

<https://doi.org/10.1145/3343031.3350937>



**Figure 2: Comparison between two unpaired translation approaches and our Mocycle-GAN.** (a) *Cycle-GAN* exploits cycle-consistency constraint to model appearance structure for unpaired image-to-image translation. (b) *Recycle-GAN* utilizes temporal predictor ( $P_X$  and  $P_Y$ ) to explore cycle consistency across both domains and time for unpaired video-to-video translation. (c) *Mocycle-GAN* explicitly models motion across frames with optical flow ( $f_{x_t}$  and  $f_{y_s}$ ), and pursuits cycle consistency on motion that enforces the reconstruction of motion. Motion translation is further exploited to transfer the motion across domains via motion translator ( $M_X$  and  $M_Y$ ), strengthening the temporal continuity in video synthesis. Dotted line denotes consistency constraint between its two endpoints.

from image-to-image to video-to-video, which is referred as an emerging problem of “unpaired video-to-video translation.” It enables a general-purpose video translation across domains in the absence of paired training data, making it flexible to be applied in a variety of video-to-video translation tasks (see Figure 1).

One straightforward way to tackle unpaired video-to-video translation is to capitalize on unpaired image-to-image translation approach, e.g., *Cycle-GAN* [37] (Figure 2(a)) that enforces an inverse translation for each frame. However, this way only explores visual appearance on frames for video synthesis and will inevitably result in temporal discontinuity when the synthetic frames are deteriorated by flickering artifacts as in video style transfer [3]. This limitation originates from the fact that video is an information-intensive media with complexities along both spatial and temporal dimensions. Such facts motivate and highlight the exploration of both appearance structure and temporal continuity in video synthesis. In this sense, not only the visual appearance in each frame but also motion between consecutive frames are ensured to be realistic and consistent for video translation.

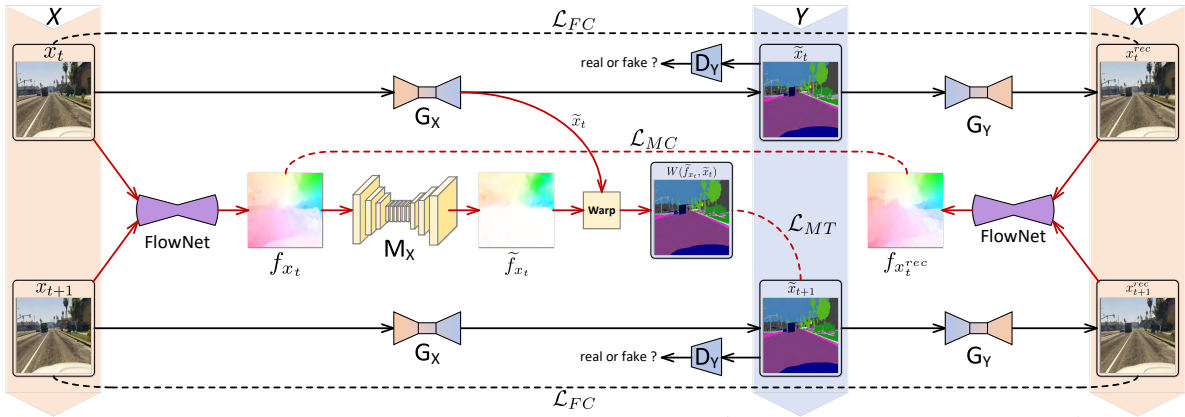
A recent pioneering practice in unpaired video-to-video translation is *Recycle-GAN* [2] (Figure 2(b)). The basic idea is to directly synthesize future frames via temporal predictor to explore cycle consistency across both domains and time. Regardless of the spatio-temporal constraint in *Recycle-GAN* for enhancing video translation, a common issue not fully studied is the exploitation of motion between consecutive frames, which is well believed to be helpful for video-to-video translation. Instead, we novelly consider the use of motion information for unpaired video-to-video translation from the viewpoint both motion cycle consistency and motion translation, as depicted in Figure 2(c). The objective of motion cycle consistency constraint is to pursuit cycle consistency on motion

between input adjacent frames, which in turn implicitly enforces the temporal continuity between synthetic adjacent frames. In addition, we exploit the constraint of motion translation to further strengthen temporal continuity in synthetic videos via transferring motion across domains. One naive method for enforcing temporal coherence is to warp the synthetic frame with the estimated motion (i.e., optical flow) between input frames to produce the subsequent frame as in [11, 28]. Nevertheless, this paradigm ignores the occlusions, blur, and appearance variations, e.g., raised by the change of lighting in different domains. As such, the temporal coherence is enforced in a brute-force manner regardless of the scene dynamics in target domain. In comparison, we leverage motion translator to transfer the estimated motion in source domain to target domain, which characterizes the temporal coherence across synthetic frames more tailored to target domain.

By consolidating the idea of exploiting motion information for facilitating unpaired video-to-video translation, we present a novel Motion-guided Cycle GAN (*Mocycle-GAN*), as shown in Figure 3. The whole architecture consists of generators and discriminators under the backbone of standard Conditional GANs, coupled with motion translator for transferring motion across domains. Specifically, the motion information in each domain is estimated in the form of optical flow between consecutive frames. During training, three types of spatial/temporal constrains, i.e., adversarial constraint, cycle consistency on both frame and motion, and motion translation, are devised to explore both the appearance structure and temporal continuity for unpaired video translation. The adversarial constraint discriminates between synthetic and real frames in an adversarial manner, making each synthetic frame realistic at appearance. For the cycle consistency on both frame and motion, it encourages the reconstruction of both appearance structure of frames and temporal continuity in motion. The motion translation constraint transfers the estimated motion from source to target domain via motion translator and then warps the synthetic frame with the transferred motion to the subsequent frame. In this sense, the temporal continuity among synthetic frames in target domain is further strengthened with the guidance from transferred motion. However, unlike in supervised video-to-video translation, we cannot train the motion translator with paired video data in unpaired scenario. Thus, we optimize the whole architecture in a Expectation Maximization (EM) procedure which iteratively updates generators and discriminators with the three spatial/temporal constrains (E-step), and refines motion translator with an auxiliary motion consistency loss (M-step). Such procedure gradually improves the motion translation as well as the video-to-video translation.

## 2 RELATED WORK

**Image-to-Image Translation.** Image-to-image translation aims to learn a mapping function from an input image in one domain to the output image in another domain. The recent advances in GANs [8] have inspired the remarkable improvement of this task [4, 13, 32, 38]. An early pioneering work [13] presents a general-purpose solution which leverages Conditional GANs for image-to-image translation. This paradigm enables a variety of graphics tasks, e.g., semantic labels to photo, edges to photo, and photo inpainting. [38] further extends [13] by encouraging the bijective consistency



**Figure 3: The overview of Mocycle-GAN for unpaired video-to-video translation ( $X$ : source domain;  $Y$ : target domain). Note that here we only depict the forward cycle  $X \rightarrow Y \rightarrow X$  for simplicity. Mocycle-GAN consists of generators ( $G_X$  and  $G_Y$ ) to synthesize frames across domains, discriminators ( $D_X$  and  $D_Y$ ) to distinguish real frames from synthetic ones, and motion translator ( $M_X$ ) for motion translation across domains. Given two real consecutive frames  $x_t$  and  $x_{t+1}$ , we firstly translate them into the synthetic frames  $\tilde{x}_t$  and  $\tilde{x}_{t+1}$  via  $G_X$ , which are further transformed into the reconstructed frames  $x_t^{rec}$  and  $x_{t+1}^{rec}$  through the inverse mapping  $G_Y$ . In addition, two optical flow  $f_{x_t}$  and  $f_{x_t^{rec}}$  are obtained by capitalizing on FlowNet to represent the motion before and after the forward cycle. During training, we leverage three kinds of spatial/temporal constrains to explore appearance structure and temporal continuity for video translation: 1) *Adversarial Constraint* ( $\mathcal{L}_{Adv}$ ) ensures each synthetic frame realistic at appearance through adversarial learning; 2) *Frame and Motion Cycle Consistency Constraint* ( $\mathcal{L}_{FC}$  and  $\mathcal{L}_{MC}$ ) encourage an inverse translation on both frames and motions; 3) *Motion Translation Constraint* ( $\mathcal{L}_{MT}$ ) validates the transfer of motion across domains in video synthesis. Specifically, the motion translator  $M_X$  converts the optical flow  $f_{x_t}$  in source to  $\tilde{f}_{x_t}$  in target, which will be utilized to further warp the synthetic frame  $\tilde{x}_t$  to the subsequent frame  $W(\tilde{f}_{x_t}, \tilde{x}_t)$ . This constraint encourages the synthetic subsequent frame  $\tilde{x}_{t+1}$  to be consistent with the warped version  $W(\tilde{f}_{x_t}, \tilde{x}_t)$  in the traceable points, leading to pixel-wise temporal continuity.**

between the latent and output spaces, leading to more realistic and diverse results. Furthermore, [15, 18, 20, 34, 35, 37] begin to tackle unsupervised image-to-image translation, i.e., learning to translate images across domains without paired data. In particular, Cycle-GAN [37] is devised to learn the mapping function in the absence of paired training data. A cycle consistency loss is utilized to train this mapping coupled with an inverse mapping between the two domains, enforcing the translation to be cycle consistent. Dual GAN [35] is a concurrent work which also exploits the cycle consistency for unpaired image-to-image translation.

Beyond the still image translation across different domains, our work pursuits its video counterpart by tackling unpaired video-to-video translation in a complex spatio-temporal context. In addition to make each frame realistic, a video translator should be capable of enhancing the temporal coherence among adjacent frames.

**Video-to-Video Translation.** Video-to-video translation is a natural extension of image-to-image translation in video domain. Specifically, [31] is one of the early attempts to tackle video-to-video translation, which integrates a spatio-temporal adversarial objective into conditional GANs. The global and local temporal consistency is exploited in [33] to ensure the local and global consistency across frames for video-to-video translation. However, the above methods require manual supervision for aligning paired videos across domains, which is extremely expensive and costly to obtain. Inspired from Cycle-GAN [37], [2] devises Recycle-GAN to facilitate unpaired video-to-video translation. Instead of solely employing spatial constraint for each frame as Cycle-GAN, Recycle-GAN additionally exploits a recurrent temporal predictor to model the dependency between nearby frames, enabling a spatio-temporal constraint (i.e., the recycle consistency) for unpaired video-to-video translation. Video style transfer is another related problem which

transfers the style of a reference image to an input video. When directly applying the image style transfer techniques [6, 7, 14, 29, 36] to videos, the generated stylized video will inevitably be affected with severe flickering artifacts. As such, to alleviate the flickering artifacts, a number of video style transfer approaches [1, 3, 5, 9, 11, 28] are proposed by additionally utilizing temporal constraints to ensure the temporal consistency across frames.

In our work, we also target for an unsupervised solution for video translation. Unlike Recycle-GAN [2] that directly predicts future frames to enforce the translation to be recycle consistent, our Mocycle-GAN explicitly models the motion across frames with optical flow and pursuits a cycle consistency on motion. Moreover, a motion translator is leveraged to transfer motion in source domain to target domain, aiming to strengthen temporal continuity across synthetic frames with the guidance from transferred motion.

### 3 APPROACH: MOCYCLE-GAN

In this paper, we devise Motion-guided Cycle GAN (Mocycle-GAN) architecture to integrate motion estimation into unpaired video translator, exploring both appearance structure and temporal continuity for video translation. The whole architecture of Mocycle-GAN is illustrated in Figure 3. We begin this section by elaborating the notation and problem formulation of unpaired video-to-video translation, followed with a brief review of Cycle-GAN with spatial constrain. Then, two kinds of motion-guided temporal constrains, i.e., motion cycle consistency and motion translation, are introduced to further strengthen the temporal continuity. In this sense, both visual appearance in each frame and motion between consecutive frames are ensured to be realistic and consistent across transformation. Finally, the optimization strategy at training along with inference stage are provided.

### 3.1 Overview

**Notation.** In unpaired video-to-video translation task, we are given two video collections:  $X = \{\mathbf{x}\}$  in source domain and  $Y = \{\mathbf{y}\}$  in target domain, where  $\mathbf{x} = \{x_t\}_{t=1}^T$  and  $\mathbf{y} = \{y_s\}_{s=1}^S$  denotes the video in source and target domain respectively.  $x_t$  and  $y_s$  represent the  $t$ -th frame in source video  $\mathbf{x}$  and  $s$ -th frame in target video  $\mathbf{y}$ . The goal of this task is to learn two mapping functions between source domain  $X$  and target domain  $Y$ , i.e.,  $G_X : X \rightarrow Y$  and  $G_Y : Y \rightarrow X$ . Here the two mapping functions  $G_X$  and  $G_Y$  are implemented as generators in Conditional GANs for synthesizing frames. As such, by performing video translation via  $G_X$  and  $G_Y$ ,  $\mathbf{x}$  and  $\mathbf{y}$  are converted as the synthetic videos  $\tilde{\mathbf{x}} = \{\tilde{x}_t\}_{t=1}^T$  and  $\tilde{\mathbf{y}} = \{\tilde{y}_s\}_{s=1}^S$ , where  $\tilde{x}_t = G_X(x_t)$  and  $\tilde{y}_s = G_Y(y_s)$  are synthetic frames. Moreover, one discriminator  $D_Y$  is leveraged to distinguish real frames  $\{y_s\}$  from synthetic ones  $\{\tilde{y}_s\}$ . Similarly, another discriminator  $D_X$  distinguishes between  $\{x_t\}$  and  $\{\tilde{x}_t\}$ . Since we ultimately aim to integrate motion estimation into video translation, we capitalize on off-the-shelf FlowNet [12] ( $\mathcal{F}$ ) to directly represent the estimated motion between two consecutive frames (e.g.,  $x_t$  and  $x_{t+1}$ ) as optical flow:  $f_{x_t} = \mathcal{F}(x_t, x_{t+1})$ . Furthermore, two motion translators, i.e.,  $M_X$  and  $M_Y$ , are devised to transfer optical flows across domains. More details about how we conduct motion translation will be elaborated in Section 3.3.

**Problem Formulation.** Inspired by the recent success of Cycle-GAN in unpaired image-to-image translation and temporal coherence/dynamics exploration in video understanding [17, 21, 22, 24], we formulate our unpaired video translation model in a cyclic paradigm which enforces the learnt mappings ( $G_X$  and  $G_Y$ ) to be cycle consistent on both frame and motion. Specifically, let  $x_t^{rec} = G_Y(G_X(x_t))$  and  $y_s^{rec} = G_X(G_Y(y_s))$  denotes the reconstructed frame of  $x_t$  and  $y_s$  in forward cycle and backward cycle, respectively. Hence the frame cycle consistency constraint aims to reconstruct each frame in source and target domain via translation cycle:  $x_t \rightarrow \tilde{x}_t \rightarrow x_t^{rec} \approx x_t$  and  $y_s \rightarrow \tilde{y}_s \rightarrow y_s^{rec} \approx y_s$ . Besides the preservation of appearance structure in translation cycle via cycle consistency on frame, we additionally pursuit the reconstruction of motion in translation cycle, which enforces the temporal continuity between consecutive frames. As such, the motion cycle consistency constraint is introduced to reconstruct the motion between every two consecutive frames through translation cycle:  $f_{x_t} \rightarrow f_{x_t^{rec}} \approx f_{x_t}$  and  $f_{y_s} \rightarrow f_{y_s^{rec}} \approx f_{y_s}$ . In addition, the motion translation constraint is especially devised to exploit motion translation across domains. The transferred motion will be directly utilized to warp the synthetic frame to the subsequent frame, which further strengthens temporal continuity among synthetic frames.

### 3.2 Cycle-GAN

We briefly review Cycle-GAN [37] for unpaired translation at frame level. Cycle-GAN is composed of two generators ( $G_X$  and  $G_Y$ ) to synthesize frames across domains, and two discriminators ( $D_X$  and  $D_Y$ ) for discriminating real frames from synthetic ones, coupled with the adversarial constraint and cycle consistency constraint on frame. The main idea behind Cycle-GAN is to make each frame realistic via adversarial constraint, and encourage the translation cycle-consistent via cycle consistency constraint on frame.

**Adversarial Constraint.** As in image/video generation [8, 23, 26, 30], the generators and discriminators are adversarially trained in a two-player minimax game mechanism. Specifically, given the real frames ( $x_t$  and  $y_s$ ) and the corresponding synthetic frames ( $\tilde{x}_t = G_X(x_t)$  and  $\tilde{y}_s = G_Y(y_s)$ ), the discriminators are trained to correctly distinguish between real and synthetic frames, i.e., maximizing the adversarial constraint:

$$\mathcal{L}_{Adv} = \sum_s \log D_Y(y_s) + \sum_t \log(1 - D_Y(\tilde{x}_t)) + \sum_t \log D_X(x_t) + \sum_s \log(1 - D_X(\tilde{y}_s)). \quad (1)$$

Meanwhile, the generators are learnt to minimize this adversarial constraint, aiming to fool the discriminators with synthetic frames.

**Frame Cycle Consistency Constraint.** Moreover, to tackle the unpaired translation, a cycle consistency constraint on each frame is additionally exploited to penalize the difference between the primary input frame  $x_t/y_s$  and its reconstructed frame  $x_t^{rec} = G_Y(G_X(x_t))/y_s^{rec} = G_X(G_Y(y_s))$ :

$$\mathcal{L}_{FC}(G_X, G_Y) = \sum_t \|x_t^{rec} - x_t\|_1 + \sum_s \|y_s^{rec} - y_s\|_1. \quad (2)$$

By minimizing the frame cycle consistency constraint above, the frame translation is enforced to be cycle-consistent, targeting to capture high-level appearance structure across domains.

### 3.3 Motion Guided Temporal Constraints

Unlike Cycle-GAN that only explores appearance structure at frame level, an unpaired video translator should further exploit temporal continuity across frames to ensure both the visual appearance and motion between frames to be realistic and consistent. Existing pioneer in unpaired video translation is Recycle-GAN [2] that predicts future frames via temporal predictor to enable the cycle consistency across both domains and time, while leaving the inherent motion information unexploited. Here we explicitly model the motion across frames in the form of optical flow throughout the translation. Two temporal constraints, i.e., motion cycle consistency and motion translation, are especially devised to strengthen temporal continuity in synthetic videos with the guidance of motion reconstruction in translation cycle and motion translation across domains.

**Motion Cycle Consistency Constraint.** To resolve unpaired scenario of video translation, we go one step further and extend the cycle consistency constraint from single frame in Cycle-GAN to motion between consecutive frames. Formally, given two consecutive frames ( $x_t$  and  $x_{t+1}$ ) from domain  $X$ , the forward translation cycle is encouraged to reconstruct the two frames ( $x_t^{rec}$  and  $x_{t+1}^{rec}$ ) with the consistent optical flow. In other words, the estimated optical flow  $f_{x_t^{rec}}$  between  $x_t^{rec}$  and  $x_{t+1}^{rec}$  should be similar to the primary optical flow  $f_{x_t}$  between  $x_t$  and  $x_{t+1}$ . Similarly, for two consecutive frames ( $y_s$  and  $y_{s+1}$ ) from domain  $Y$ , the backward translation cycle is enforced to be cycle-consistent on optical flow:  $f_{y_s} \rightarrow f_{y_s^{rec}} \approx f_{y_s}$ . Accordingly, the motion cycle consistency constraint is defined as the  $L_1$  distance between the optical flows before and after the translation cycle:

$$\mathcal{L}_{MC}(G_X, G_Y) = \sum_t \sum_i C_{x_t}^{(i)} \left\| f_{x_t^{rec}}^{(i)} - f_{x_t}^{(i)} \right\|_1 + \sum_s \sum_i C_{y_s}^{(i)} \left\| f_{y_s^{rec}}^{(i)} - f_{y_s}^{(i)} \right\|_1, \quad (3)$$

where  $f_{x_t}^{(i)}$  denotes a 2-dimensional displacement vector for  $i$ -th pixel in optical flow  $f_{x_t}$ . As in [11], we leverage two visibility masks  $C_{x_t}$  and  $C_{y_s}$  as weight matrixes, where each pixel  $C_{x_t}^{(i)}, C_{y_s}^{(i)} \in [0, 1]$  represents the per-pixel confidence of the pixel  $f_{x_t}^{(i)}$  in optical flow  $f_{x_t}$ : 1 for traceable pixels by optical flow, and 0 at occluded regions or near motion boundaries. Accordingly, by minimizing the motion cycle consistency constraint, the video translation is ensured to preserve the motion between real consecutive frames after translation cycle, which in turn implicitly enhances the temporal continuity between synthetic consecutive frames.

**Motion Translation Constraint.** The cycle consistency on motion only constraints temporal coherence between synthetic frames in an unsupervised manner, but ignores the straightforward transfer of motion across domains. Nevertheless, the transfer of motion across domains has been seldom exploited for unpaired video translation, possibly because such motion translation needs pairs of optical flows for training, while in the unpaired settings, no paired video data is provided. One naive way to exploit motion across domains for video synthesis is to directly warp the synthetic frame with the source motion into the subsequent frame as in [11, 28]. This scheme pursuits the motion consistency across domains in a brute-force manner regardless of the scene dynamics in target. Instead, we design a novel motion translator to transfer optical flow from source domain to target domain, which captures temporal coherence tailored to target domain. Such transferred optical flow via motion translator can be further leveraged to guide video synthesis in target domain, pursuing the pixel-wise temporal continuity.

Technically, given the optical flow  $f_{x_t}$  between  $x_t$  and  $x_{t+1}$  from domain  $X$ , the motion translator  $M_X$  is utilized to transform the primary optical flow  $f_{x_t}$  into the transferred one  $\tilde{f}_{x_t} = M_X(f_{x_t})$  in domain  $Y$ . Note that motion translators are implemented as paired translator Pix2Pix in [13]. Each motion translator is constrained with an auxiliary motion consistency loss, aiming to correctly predict the optical flow in the target domain. Here we directly utilize the optical flow  $f_{\tilde{x}_t}$  between the corresponding synthetic frames in target domain as the ‘‘pseudo’’ target optical flow for training motion translator. Similarly, with the input of optical flow  $f_{y_s}$  from domain  $Y$ , another motion translator  $M_Y$  produces the transferred optical flow  $\tilde{f}_{y_s} = M_Y(f_{y_s})$  in domain  $X$ , which is enforced to resemble the ‘‘pseudo’’ target optical flow  $f_{\tilde{y}_s}$  in domain  $X$ . Thus, the auxiliary motion consistency loss is defined as  $L_1$  distance between the transferred optical flow and ‘‘pseudo’’ target optical flow:

$$\mathcal{L}_{AM}(M_X, M_Y) = \sum_t \left\| \tilde{f}_{x_t} - f_{\tilde{x}_t} \right\|_1 + \sum_s \left\| \tilde{f}_{y_s} - f_{\tilde{y}_s} \right\|_1. \quad (4)$$

After that, the transferred optical flow  $\tilde{f}_{x_t}/\tilde{f}_{y_s}$  is utilized to further warp the synthetic frame  $\tilde{x}_t/\tilde{y}_s$  to the subsequent frame via bi-linear interpolation, leading to the warped frame  $W(\tilde{f}_{x_t}, \tilde{x}_t)/W(\tilde{f}_{y_s}, \tilde{y}_s)$  in target domain at time  $t + 1$ . Therefore, we define the motion translation constraint as the  $L_1$  distance between the warped frame and the synthetic frame at time  $t + 1$ :

$$\mathcal{L}_{MT}(G_X, G_Y) = \sum_t \sum_i C_{x_t}^{(i)} \left\| W^{(i)}(\tilde{f}_{x_t}, \tilde{x}_t) - \tilde{x}_{t+1}^{(i)} \right\|_1 + \sum_s \sum_i C_{y_s}^{(i)} \left\| W^{(i)}(\tilde{f}_{y_s}, \tilde{y}_s) - \tilde{y}_{s+1}^{(i)} \right\|_1. \quad (5)$$

This motion translation constraint ensures the synthetic frame to be consistent with the warped version of previous synthetic frame

---

### Algorithm 1 The training process of Mocycle-GAN

---

```

1: Input: The number of maximum training iteration  $N$ ; Initialize generators ( $G_X, G_Y$ ), discriminators ( $D_X, D_Y$ ), and motion translators ( $M_X, M_Y$ ).
2: for  $n = 1$  to  $N$  do
3:   Fetch input batch with sampled consecutive frame pairs  $\{(x_t, x_{t+1}), (y_s, y_{s+1})\}$ .
4:   for Each consecutive frame pair  $(x_t, x_{t+1}), (y_s, y_{s+1})$  do
5:     Generate synthetic frames  $(\tilde{x}_t, \tilde{x}_{t+1}), (\tilde{y}_s, \tilde{y}_{s+1})$  and reconstructed frames  $(x_t^{rec}, x_{t+1}^{rec}), (y_s^{rec}, y_{s+1}^{rec})$  via generators ( $G_X, G_Y$ ).
6:     Calculate the corresponding optical flow  $f_{x_t}, f_{y_s}, f_{\tilde{x}_t}, f_{\tilde{y}_s}, f_{x_t^{rec}}, f_{y_s^{rec}}$  via FlowNet.
7:     Produce the transferred flow  $\tilde{f}_{x_t}$  and  $\tilde{f}_{y_s}$  via motion translators ( $M_X, M_Y$ ).
8:   end for
9:   E-step:
10:    Fix motion translators ( $M_X, M_Y$ ).
11:    Update generators ( $G_X, G_Y$ ) w.r.t loss in Eq.(6).
12:    Update discriminators ( $D_X, D_Y$ ) w.r.t loss in Eq.(1).
13:   M-step:
14:    Fix generators ( $G_X, G_Y$ ) and discriminators ( $D_X, D_Y$ ).
15:    Update motion translators ( $M_X, M_Y$ ) w.r.t loss in Eq.(4).
16: end for

```

---

in the traceable points. As such, the pixel-wise temporal continuity among synthetic frames are strengthened.

## 3.4 Training and Inference

**Optimization.** The overall training objective of our Mocycle-GAN integrates the adversarial constraint, the cycle consistency constraints on frame and motion, and the motion translation constraint for generators and discriminators, and the auxiliary motion consistency loss for motion translators. During training, we adopt the EM procedure to iteratively optimize motion translators, and generators & discriminators. Specifically, in **E-step**, we fix the parameters in motion translators ( $M_X$  and  $M_Y$ ) and update the parameters of generators ( $G_X$  and  $G_Y$ ) by minimizing the combined loss of the three spatial/temporal constrains:

$$\mathcal{L}(G_X, G_Y) = \mathcal{L}_{Adv} + \lambda_{FC} \cdot \mathcal{L}_{FC}(G_X, G_Y) + \lambda_{MC} \cdot \mathcal{L}_{MC}(G_X, G_Y) + \lambda_{MT} \cdot \mathcal{L}_{MT}(G_X, G_Y), \quad (6)$$

where  $\lambda_{FC}$ ,  $\lambda_{MC}$ , and  $\lambda_{MT}$  are tradeoff parameters. Meanwhile, the discriminators ( $D_X$  and  $D_Y$ ) are optimized by maximizing the adversarial constraint  $\mathcal{L}_{Adv}$  in Eq.(1). In **M-step**, we fix the parameters in generators and discriminators, and update motion translators by minimizing the auxiliary motion consistency loss  $\mathcal{L}_{AM}(M_X, M_Y)$  in Eq.(4). We alternate the E-step and M-step in each training iteration until a convergence criterion is met. The detailed training process of our Mocycle-GAN is given in Algorithm 1. Note that in practice, the generators & discriminators are pre-trained with the combined loss of adversarial constraint and cycle consistency constraints on frame & motion. Next, we pre-train the motion translators with the auxiliary motion consistency loss.

**Inference.** After the optimization of our Mocycle-GAN, we can obtain the learnt generator  $G_X$  and motion translator  $M_X$ . During inference, given an input video  $\mathbf{x} = \{x_t\}_{t=1}^T$ , the simplest way for video translation is to directly employ generator  $G_X$  to convert  $\mathbf{x}$  into the synthetic video  $\tilde{\mathbf{x}} = \{\tilde{x}_t\}_{t=1}^T$  frame-by-frame. An alternative solution is to leverage the warped version of previous synthetic frame based on the transferred optical flow to smooth the output:

$$\tilde{x}_{t+1} = \frac{G_X(x_{t+1}) + W(\tilde{f}_{x_t}, \tilde{x}_t)}{2}. \quad (7)$$

However, for fair comparison to other image/video translation approaches, we adopt the simplest single-frame translation without any post processing for evaluation in the experiments.

**Table 1: Segmentation score (%) of our Mocycle-GAN and other methods for video-to-labels translation on Viper.**

Criterion	Approach	day	sunset	rain	snow	night	all
<b>MP</b>	Cycle-GAN [37]	46.0	68.7	41.1	39.2	32.2	40.2
	Recycle-GAN [2]	53.0	75.6	51.6	55.5	39.7	58.8
	Recycle-GAN <sub>cmb</sub> [2]	54.7	76.3	51.0	57.0	44.7	60.1
	Cycle-GAN <sub>SF</sub> [11]	55.2	77.1	49.9	59.6	42.2	62.3
	Mocycle-GAN	<b>64.2</b>	<b>82.1</b>	<b>67.0</b>	<b>66.1</b>	<b>64.5</b>	<b>64.9</b>
<b>AC</b>	Cycle-GAN [37]	12.0	13.1	5.1	9.5	4.9	9.6
	Recycle-GAN [2]	13.5	16.8	9.9	11.0	8.4	14.4
	Recycle-GAN <sub>cmb</sub> [2]	15.3	15.7	10.9	11.3	10.2	14.9
	Cycle-GAN <sub>SF</sub> [11]	16.2	17.0	10.7	13.0	9.7	16.1
	Mocycle-GAN	<b>20.5</b>	<b>23.0</b>	<b>18.4</b>	<b>17.8</b>	<b>16.4</b>	<b>17.7</b>
<b>IoU</b>	Cycle-GAN [37]	7.4	9.9	3.1	5.8	2.9	5.3
	Recycle-GAN [2]	9.4	13.1	6.6	7.8	5.2	10.5
	Recycle-GAN <sub>cmb</sub> [2]	10.8	12.4	6.8	8.1	6.4	11.0
	Cycle-GAN <sub>SF</sub> [11]	11.6	13.4	6.3	9.0	6.4	11.0
	Mocycle-GAN	<b>15.2</b>	<b>18.1</b>	<b>11.9</b>	<b>12.3</b>	<b>11.6</b>	<b>13.2</b>

## 4 EXPERIMENTS

We empirically verify the merit of our Mocycle-GAN by conducting experiments on four different unpaired video translation scenarios, including video-to-labels, labels-to-video, four ambient condition transfers (day-to-night, night-to-day, day-to-sunset, sunset-to-day) on Viper [27] and flower-to-flower on Flower Video Dataset [2].

### 4.1 Datasets and Experimental Settings

**Viper** is a popular visual perception benchmark to facilitate both low-level and high-level vision tasks, e.g., optical flow and semantic segmentation. It consists of videos from a realistic virtual world (i.e. GTA gameplay), which are collected while driving, riding and walking in diverse ambient conditions (day, sunset, snow, rain, and night). Each frame (resolution:  $1920 \times 1080$ ) is annotated with pix-level labels, i.e., segmentation label map. Following [2], we split 77 videos under diverse environmental conditions into 57 for training and 20 for testing. For video-to-labels and labels-to-video, we evaluate the translations between videos and segmentation label maps. For ambient condition transfers, we consider the translation across different ambient conditions: day  $\leftrightarrow$  night and day  $\leftrightarrow$  sunset.

**Flower Video Dataset** is a recent released dataset for video translation. This dataset includes the time-lapse videos which depict the blooming or fading of various flowers but without any sync. The resolution of each video is  $256 \times 256$ . For flower-to-flower, we evaluate the translation between different types of flowers, aiming to align the high-level semantic content among them, e.g., the two flowers simultaneously bloom or fade at the same pace.

**Implementation Details.** We mainly implement our Mocycle-GAN on Pytorch [25] architecture. For generators, we follow the settings of [2, 37], and adopt the encoder-decoder architecture [14]. In particular, each generator is composed of two convolution layers (stride: 2) for down-sampling, six residual blocks [10], and two deconvolution layers for up-sampling. Each discriminator is built as the  $70 \times 70$  PatchGAN in [13]. For motion translators, we adopt the similar architecture of generator by modifying the input and output channel as 2, which enables the translation of optical flow across domains. In all experiments, we set the tradeoff parameters in Eq.(6) as  $\lambda_{FC} = 10$ ,  $\lambda_{MC} = 10$ , and  $\lambda_{MT} = 10$ . During training, the batch size is set as 1. Adam [16] is utilized to optimize the parameters in generators, discriminators and motion translators with the initial learning rate of 0.0002, 0.0002 and 0.0001, respectively.

**Table 2: FCN score (%) of our Mocycle-GAN and other methods for labels-to-video translation on Viper.**

Criterion	Approach	day	sunset	rain	snow	night	all
<b>MP</b>	Cycle-GAN [37]	36.3	48.7	23.7	40.0	22.8	37.9
	Recycle-GAN [2]	37.5	53.9	27.4	42.7	23.6	41.3
	Recycle-GAN <sub>cmb</sub> [2]	37.0	54.4	27.6	40.8	26.6	43.5
	Cycle-GAN <sub>SF</sub> [11]	38.7	57.0	25.2	42.1	24.4	44.6
	Mocycle-GAN	<b>42.1</b>	<b>61.2</b>	<b>34.6</b>	<b>48.1</b>	<b>30.5</b>	<b>47.6</b>
<b>AC</b>	Cycle-GAN [37]	10.7	15.3	9.1	11.4	10.0	10.2
	Recycle-GAN [2]	12.3	14.9	10.0	11.5	11.1	12.0
	Recycle-GAN <sub>cmb</sub> [2]	12.7	15.6	10.1	12.0	11.8	12.2
	Cycle-GAN <sub>SF</sub> [11]	13.2	15.4	9.7	13.0	10.4	12.9
	Mocycle-GAN	<b>15.4</b>	<b>17.6</b>	<b>12.6</b>	<b>14.9</b>	<b>16.5</b>	<b>16.0</b>
<b>IoU</b>	Cycle-GAN[37]	7.4	9.2	4.7	6.2	4.5	6.1
	Recycle-GAN [2]	8.1	10.0	5.5	6.9	4.7	6.7
	Recycle-GAN <sub>cmb</sub> [2]	8.3	10.2	5.5	6.9	5.6	7.0
	Cycle-GAN <sub>SF</sub> [11]	8.0	10.4	5.0	7.0	5.3	7.4
	Mocycle-GAN	<b>9.7</b>	<b>11.9</b>	<b>7.5</b>	<b>8.8</b>	<b>7.7</b>	<b>10.1</b>

**Evaluation Metrics.** For video-to-labels translation, as in [2, 37], we adopt three standard segmentation metrics in [19] for evaluation, i.e., Mean Pixel Accuracy (**MP**), Average Class Accuracy (**AC**), and Intersection-Over-Union (**IoU**). For labels-to-video translation, we follow [2, 37] and report the **FCN score** on target domain. FCN score represents the quality of synthetic frames according to an off-the-shelf semantic segmentation network. Specifically, we pre-train a fully-convolutional network, i.e., FCN [19], on Viper. Next, the FCN model is utilized to predict the segmentation label map for each synthetic frame. By comparing the predicted segmentation label map against the ground-truth labels, we can obtain the FCN scores with regard to the three standard segmentation metrics described above (i.e., MP, AC, and IoU). The intuition is that the higher the FCN scores, the more realistic the synthetic frames at appearance.

**Compared Approaches.** We include the following state-of-the-art unpaired translation methods for performance comparison: (1) **Cycle-GAN**[37] is an unpaired image translator that pursues an inverse translation only at frame level. (2) **Recycle-GAN**[2] leverages a recurrent temporal predictor to generate future frames and pursues a new cycle consistency (i.e. recycle loss) across domains and time for unpaired video-to-video translation. (3) **Recycle-GAN<sub>cmb</sub>** [2] is an upgraded version of Recycle-GAN by combining recycle loss in Recycle-GAN and cycle loss in Cycle-GAN for training video translator. (4) **Cycle-GAN<sub>SF</sub>** remoulds a state-of-the-art video style transfer approach [11] for unpaired video translation by equipping its short temporal constraint with the cycle loss in Cycle-GAN. The basic idea of the short temporal constraint is to directly warp the synthetic frame with the source motion into the subsequent frame, aiming to enforce the pixel-wise temporal consistency. (4) **Mocycle-GAN** is the proposal in this paper. Please note that for fair comparison, all the baselines and our Mocycle-GAN utilize the same architecture for generators and discriminators.

### 4.2 Performance Comparison and Analysis

**Evaluation on Video-to-Labels.** In this scenario, the video translator takes a game scene video as input and outputs the corresponding segmentation label maps. The performance comparisons of different models for video-to-labels translation task are summarized in Table 1. Overall, the results across three segmentation metrics consistently indicate that our proposed Mocycle-GAN obtains better performances against state-of-the-art techniques. The results



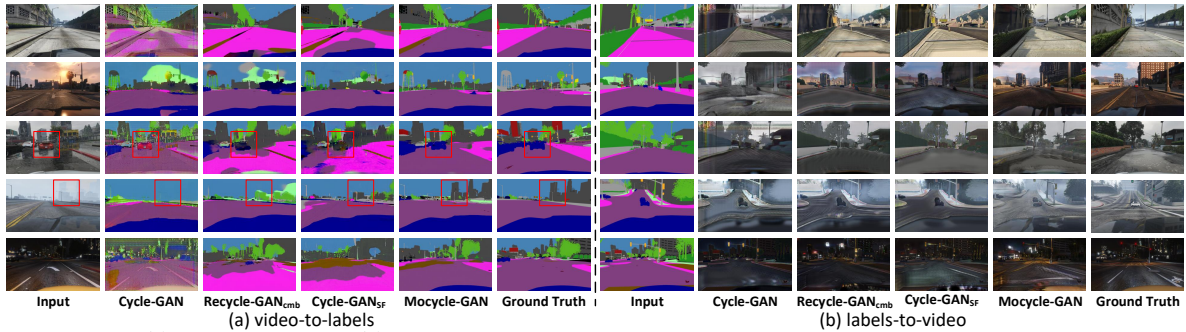


Figure 4: Examples of (a) video-to-labels and (b) labels-to-video results in Viper dataset under various ambient conditions. The original inputs, the output results by different models, and the ground truth outputs are given.

Table 3: Ablation study for each design (i.e., Motion Cycle Consistency (MC) and Motion Translation (MT)) in Mocycle-GAN for video-to-labels on Viper.

Criterion	Approach	MC	MT	day	sunset	rain	snow	night	all
MP	Cycle-GAN + MC	√		60.2	81.1	61.3	63.0	50.7	63.1
	Cycle-GAN + MT		√	62.5	81.5	65.4	64.6	63.0	63.0
	Mocycle-GAN	√	√	<b>64.2</b>	<b>82.1</b>	<b>67.0</b>	<b>66.1</b>	<b>64.5</b>	<b>64.9</b>
AC	Cycle-GAN + MC	√		18.2	21.3	15.7	14.4	12.2	17.4
	Cycle-GAN + MT		√	19.3	21.4	17.6	17.4	16.1	17.2
	Mocycle-GAN	√	√	<b>20.5</b>	<b>23.0</b>	<b>18.4</b>	<b>17.8</b>	<b>16.4</b>	<b>17.7</b>
IoU	Cycle-GAN + MC	√		13.3	16.9	9.6	10.4	7.9	12.9
	Cycle-GAN + MT		√	14.4	17.1	11.5	11.9	11.1	12.8
	Mocycle-GAN	√	√	<b>15.2</b>	<b>18.1</b>	<b>11.9</b>	<b>12.3</b>	<b>11.6</b>	<b>13.2</b>

generally highlight the key advantage of exploring motion information for unpaired video translation, enforcing the synthetic videos to be both realistic at appearance and temporal continuous across frames. Specifically, by encouraging the cycle consistency across domains and time via a spatio-temporal constraint, Recycle-GAN exhibits better performance than Cycle-GAN that only pursues cycle consistency at frame level. Moreover, by simultaneously utilizing the spatial constraint in Cycle-GAN and spatio-temporal constraint in Recycle-GAN, Recycle-GAN<sub>cmb</sub> further boosts up the performances. Different from Recycle-GAN<sub>cmb</sub> that enforces temporal coherence via future frame prediction, Cycle-GAN<sub>SF</sub> encourages pixel-wise temporal consistency by directly warping the synthetic frame with source optical flow, and achieves better performances. This confirms the effectiveness of modeling motion information in video synthesis. Nevertheless, the performances of Cycle-GAN<sub>SF</sub> are still lower than our Mocycle-GAN which further strengthens temporal continuity via motion cycle consistency and motion translation across domains.

Figure 4(a) showcases five examples of video-to-labels results with different methods under various ambient conditions. As illustrated in the figure, our Mocycle-GAN obtains much more promising video-to-labels results. For instance, the majority categories, e.g., road (first row), cannot be well translated for baselines. Instead, even the minority classes such as car (third row) and building (fourth row) are translated nicely using our Mocycle-GAN.

**Evaluation on Labels-to-Video.** In this scenario, given an input sequence of segmentation label maps, the video translator outputs a video that resembles a real game scene video. Table 2 shows the results on labels-to-video translation task on Viper. Our Mocycle-GAN performs consistently better than other methods over three metrics. Similar to the observations on the video-to-labels translation task, Recycle-GAN exhibits better performance

Table 4: Ablation study for each design (i.e., Motion Cycle Consistency (MC) and Motion Translation (MT)) in Mocycle-GAN for labels-to-video on Viper.

Criterion	Approach	MC	MT	day	sunset	rain	snow	night	all
MP	Cycle-GAN + MC	√		40.3	58.6	29.5	43.8	27.9	44.7
	Cycle-GAN + MT		√	39.0	57.7	33.3	46.3	27.7	47.0
	Mocycle-GAN	√	√	<b>42.1</b>	<b>61.2</b>	<b>34.6</b>	<b>48.1</b>	<b>30.5</b>	<b>47.6</b>
AC	Cycle-GAN + MC	√		14.5	16.3	11.0	13.2	14.7	13.6
	Cycle-GAN + MT		√	14.6	16.1	11.3	13.9	14.5	14.5
	Mocycle-GAN	√	√	<b>15.4</b>	<b>17.6</b>	<b>12.6</b>	<b>14.9</b>	<b>16.5</b>	<b>16.0</b>
IoU	Cycle-GAN + MC	√		9.4	11.0	6.2	7.3	7.0	7.6
	Cycle-GAN + MT		√	9.2	11.0	6.5	8.2	6.7	8.6
	Mocycle-GAN	√	√	<b>9.7</b>	<b>11.9</b>	<b>7.5</b>	<b>8.8</b>	<b>7.7</b>	<b>10.1</b>

than Cycle-GAN, by synthesising future frames via temporal predictor to explore cycle consistency across both domains and time. The further performance improvement is attained when combining Cycle-GAN and Recycle-GAN. In addition, Cycle-GAN<sub>SF</sub> explores motion across domains to directly constrain the temporal dynamics between synthetic frames with source motion and achieves better performances than Recycle-GAN<sub>cmb</sub>. Furthermore, by steering unpaired video translation with the guidance from motion cycle consistency and motion translation across domains, our Mocycle-GAN boosts up the performances over all the three metrics.

Figure 4(b) shows five examples of labels-to-video results under variant ambient conditions. Clearly, our Mocycle-GAN generates more natural and vivid frames compared with the results of baselines. Concretely, our results contain more realistic objects (e.g., road, tree, and car) with plenty of details, while the other methods always generate repeated patterns and fail to capture the details.

**Ablation Study.** In this section, we further study how each design in our Mocycle-GAN affects the overall performance. Motion Cycle consistency (MC) exploits the cycle consistency on motion to enforce the reconstruction of motion through translation cycle. Motion Translation (MT) transfers the optical flow across domains and further strengthens the temporal continuity in target domain by steering video translation with the transferred optical flow. Table 3 and Table 4 details the performance improvements by considering different designs for video-to-labels and labels-to-video on Viper, respectively. In particular, by further integrating motion cycle consistency and motion translation constraint into Cycle-GAN, Cycle-GAN + MC and Cycle-GAN + MT exhibit better performance than Cycle-GAN. Combining the two motion-guided temporal constraints, our Mocycle-GAN obtains the best performances on both video-to-labels and labels-to-video translations.



Figure 5: Examples of night-to-day results in Viper dataset. The original inputs and the output results by different models are given. Each row denotes one sequence of frames.



Figure 6: Examples of motion translation results in video-to-labels. From left to right: Source frame overlay, optical flow in source, transferred optical flow via motion translator, ground truth optical flow in target, and ground truth target frame overlay.

Moreover, to fully verify the effectiveness of the devised motion translation constraint, here we compare Cycle-GAN + MT against the best competitor Cycle-GAN<sub>SF</sub> which also exploits the motion information across domains. Unlike Cycle-GAN<sub>SF</sub> enforces the temporal coherence among synthetic frames in a brute-force manner, Cycle-GAN + MT elegantly transfers optical flow across domains to model the temporal coherence in target domain and thus achieves better performances. Figure 6 further showcases two examples of motion translation in video-to-labels. As illustrated in the figure, the optical flows in source and target domains are substantially different, and the transferred optical flow obtained by our motion translator ends up matching closely to the ground truth optical flow in target. The results again confirms the importance of transferring motion across domains for video translation.

### 4.3 Other Video Translations

**Ambient Condition transfer.** As an universal unpaired video translator, we test our Mocycle-GAN on ambient condition transfers which explore the translation between different ambient conditions. Figure 5 shows the translated videos by our Mocycle-GAN and other baselines on night-to-day task. As depicted in the figure, the baselines all generate frames whose overall color is somewhat bleak. In contrast, the color of our results gets much brighter, which better matches the style of day-time videos. Besides, our Mocycle-GAN takes the advantages of exploring both motion cycle consistency and motion translation, and thus achieves more realistic and temporal consistent videos than other methods.

**Flower-to-Flower.** We further evaluate our Mocycle-GAN on flower-to-flower that considers the translation between different flowers. The examples of translated videos by different methods are shown in Figure 7. Similar to the observations for ambient condition transfer, our Mocycle-GAN generates the most realistic and temporal continuous frames, where the target flower blooms



Figure 7: Examples of flower-to-flower results. The original inputs and the output results by different models are given. Each row denotes one sequence of frames.

Table 5: Human preference score (%) on translation quality for ambient condition transfer and flower-to-flower.

Human preference score	Ambient Condition Transfer	Flower-to-Flower
Mocycle-GAN / Cycle-GAN	82.5 / 17.5	77.5 / 22.5
Mocycle-GAN / Recycle-GAN <sub>cmb</sub>	73.8 / 26.2	72.5 / 27.5
Mocycle-GAN / Cycle-GAN <sub>SF</sub>	66.3 / 33.7	88.8 / 11.2

and fades in synch with the source flower. This again validates the effectiveness of guiding video translation with motion information.

**Human Evaluation.** We additionally conducted a human study to quantitatively evaluate Mocycle-GAN against three baselines, i.e., Cycle-GAN, Recycle-GAN<sub>cmb</sub>, and Cycle-GAN<sub>SF</sub> on ambient condition transfer and flower-to-flower tasks. For each task, we invite 10 labelers and randomly select 80 videos clips from testing set for human evaluation. We show each input video clip with two translated results (generated by our Mocycle-GAN and one baseline) at a time and ask the labelers: which one looks more realistic and natural? According to all labelers’ feedback, we measure the human preference score of one method as the percentage of its translation results that are preferred. Table 5 shows the results of human study. Clearly, our Mocycle-GAN is the winner on both translation tasks.

## 5 CONCLUSIONS

We have presented Motion-guided Cycle GAN (Mocycle-GAN) architecture, which explores both appearance structure and temporal continuity for video-to-video translation in an unsupervised manner. In particular, we study the problem from the viewpoint of integrating motion estimation into unpaired video translator. To verify our claim, we devise three types of spatial/temporal constraints: adversarial constraint is to discriminate between synthetic and real frames in an adversarial manner and thus enforce each synthetic frame realistic at appearance; frame and motion cycle consistency constraints encourage the reconstruction of both appearance structure in frames and temporal continuity in motion; motion translation constraint validates the transfer of motion across domains which further strengthens the temporal continuity. Extensive experiments conducted on video-to-labels and labels-to-video translation validate our proposal and analysis. More remarkably, the qualitative results and human study on more translations, e.g., flower-to-flower and ambient condition transfer, demonstrate the efficacy of Mocycle-GAN.

**Acknowledgments.** This work was supported in part by NSFC projects 61872329 and 61572451.



## REFERENCES

- [1] Alexander G Anderson, Cory P Berg, Daniel P Mossing, and Bruno A Olshausen. 2016. Deepmovie: Using optical flow and deep neural networks to stylize movies. *arXiv preprint arXiv:1605.08153* (2016).
- [2] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-gan: Unsupervised video retargeting. In *ECCV*.
- [3] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. In *ICCV*.
- [4] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.
- [5] Chang Gao, Derun Gu, Fangjun Zhang, and Yizhou Yu. 2018. ReCoNet: Real-time Coherent Video Style Transfer Network. *arXiv preprint arXiv:1807.01197* (2018).
- [6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *CVPR*.
- [7] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *BMVC*.
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.
- [9] Agrim Gupta, Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2017. Characterizing and improving stability in neural style transfer. In *ICCV*.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [11] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. Real-time neural style transfer for videos. In *CVPR*.
- [12] Eddy Ilg, Nikolaus Mayer, Tommo Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*.
- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.
- [14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.
- [15] Taeksoo Kim, Moon-su Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.
- [16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.
- [17] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *CVPR*.
- [18] Ming-Yu Liu and Once Tuzel. 2016. Coupled generative adversarial networks. In *NIPS*.
- [19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.
- [20] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. 2018. Unsupervised Attention-guided Image-to-Image Translation. In *NIPS*.
- [21] Yingwei Pan, Yehao Li, Ting Yao, Tao Mei, Houqiang Li, and Yong Rui. 2016. Learning Deep Intrinsic Video Representation by Exploring Temporal Coherence and Graph Structure. In *IJCAI*.
- [22] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *CVPR*.
- [23] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *ACM MM*.
- [24] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *CVPR*.
- [25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, et al. 2017. Automatic differentiation in PyTorch. In *NIPS Workshop*.
- [26] Zhaofan Qiu, Yingwei Pan, Ting Yao, and Tao Mei. 2017. Deep semantic hashing with generative adversarial networks. In *SIGIR*.
- [27] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for benchmarks. In *ICCV*.
- [28] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *GCPR*.
- [29] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *ICML*.
- [30] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *NIPS*.
- [31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *NIPS*.
- [32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*.
- [33] Xingxing Wei, Jun Zhu, Sitong Feng, and Hang Su. 2018. Video-to-video translation with global temporal consistency. In *ACM MM*.
- [34] Xuewen Yang, Dongliang Xie, and Xin Wang. 2018. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. In *ACM MM*.
- [35] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.
- [36] Hang Zhang and Kristin Dana. 2018. Multi-style generative network for real-time transfer. In *ECCV*.
- [37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
- [38] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *NIPS*.