Mutual-GAN: Towards Unsupervised Cross-Weather Adaptation with Mutual Information Constraint

Jiawei Chen, Yuexiang Li, Kai Ma, Yefeng Zheng
Tencent Jarvis Lab, Shenzhen, China
vicyxli@tencent.com

Abstract

Convolutional neural network (CNN) have proven its success for semantic segmentation, which is a core task of emerging industrial applications such as autonomous driving. However, most progress in semantic segmentation of urban scenes is reported on standard scenarios, i.e., daytime scenes with favorable illumination conditions. In practical applications, the outdoor weather and illumination are changeable, e.g., cloudy and nighttime, which results in a significant drop of semantic segmentation accuracy of CNN only trained with daytime data. In this paper, we propose a novel generative adversarial network (namely Mutual-GAN) to alleviate the accuracy decline when daytime-trained neural network is applied to videos captured under adverse weather conditions. The proposed Mutual-GAN adopts mutual information constraint to preserve image-objects during cross-weather adaptation, which is an unsolved problem for most unsupervised image-to-image translation approaches (e.g., CycleGAN). The proposed Mutual-GAN is evaluated on two publicly available driving video datasets (i.e., CamVid and SYNTHIA). The experimental results demonstrate that our Mutual-GAN can yield visually plausible translated images and significantly improve the semantic segmentation accuracy of daytime-trained deep learning network while processing videos under challenging weathers.

1. Introduction

As one of the key techniques for autonomous driving, semantic segmentation of urban scenes gains increasing attentions from the community. Due to the development of convolutional neural networks (CNNs), semantic segmentation shows remarkable efficiency and reliability in the standard scenarios, e.g., daytime scenes with favorable illumination conditions. Yet, many existing systems have not been applied to practical applications, since they cannot deal with the adverse weathers such as nighttime and cloudy due to the lack of training data. Labeling more data under adverse weathers for network training may be one of the solutions to improve the semantic segmentation accuracy of CNNs. However, the pixel-wise annotation for semantic segmentation is laborious. Additionally, the illuminance of adverse conditions is often extremely weak that drastically deteriorates the structure, texture and color features of objects (e.g., road, car and pedestrian), which significantly increases the difficulty of accurate manual annotation.

One of the promising solutions is translating the images under adverse conditions to the standard scene (i.e., daytime) for accurate annotation/segmentation. Generative adversarial network (GAN) [14], which recently achieves a great success on generating high-quality synthetic images, becomes the mainstream for image translation. Many GAN-based image-to-image (I2I) translation approaches have been proposed, e.g., CycleGAN [33], UNIT [20], and DRIT [18]. However, these GANs are frequently observed to yield the corruptions of image content in the translated images, which is unacceptable for the cross-weather translation task, due to the requirement of rigorous preservation of image-objects (e.g., car and pedestrian). To address the problem, Huang et al. [15] employed an additional segmentation to embed the semantic information to the generators, which enforced the CycleGAN to preserve the image-
objects during daytime-nighttime translation. Nevertheless, the obvious drawback of this method is the demand of pixel-wise annotation of daytime and nighttime images.

In this paper, we propose a novel GAN, namely Mutual-GAN, for the cross-weather translation of urban scenes. A mutual information constraint, approximating the lower bound of the mutual information between source and translated images, is proposed to prevent image-objects from corruptions, which is not yet addressed by most GAN-based I2I approaches. Two publicly available driving video datasets (i.e., CamVid and SYNTHIA) are used to evaluate the effectiveness of our Mutual-GAN. Experimental results show that the proposed Mutual-GAN can not only produce plausible cross-weather images, while impeccably preserving the image-objects, but also significantly increase the semantic segmentation accuracy of daytime-trained networks (e.g., PSPNet [32]) under adverse conditions.

2. Related Work

The areas highly related to our study are unsupervised I2I translation and mutual information. In this section, we briefly review the previous works on these topics.

2.1. Unsupervised Image-to-Image Translation

Witnessing the success of cycle-consistency-based approaches (e.g., CycleGAN [33], DiscoGAN [16], and DualGAN [29]), an increasing number of researchers [25] [7] [21] [22] made their efforts to the area of unpaired I2I translation. For example, UNIT [20], a recently proposed model, assumes that there exists a shared-latent space in which a pair of corresponding images from different domains could be mapped to the same latent representation. Through such latent representation, the I2I translation can be achieved. To further increase the output diversity, Lee et al. [18] proposed a disentangled representation framework, namely DRIT, with unpaired training data. DRIT embedded images into two spaces—a domain-invariant content space capturing shared information across domains, and a domain-specific attribute space to achieve diversity of the translated results. However, none of those approaches explicitly takes the image content preservation into account during translation, which may result in content distortion of the translated images.

2.2. Domain Adaptation of Urban Scene

Generative adversarial networks have been widely used for the translation of urban scenes such as daytime-nighttime [11] [15] [26], synthetic-realistic [8] [9] [23], and cross-dataset [10] [19] adaptation. In this paper, we focus on the cross-weather adaptation, i.e., translation between the standard (daytime) and adverse (nighttime and cloudy) scenes. A typical study in this area is AugGAN [15], which performed satisfactory daytime-nighttime translation to augment the amount of training data for neural networks. Nonetheless, AugGAN required pixel-wise annotations to overcome the aforementioned problem (i.e., image-object corruptions in the translated images), which were difficult to acquire and inapplicable for practical applications.

To reduce the workload of manual annotation, we try to address the problem of content distortion in an unsupervised manner. Chen et al. [5] implemented a VideoGAN to maintain the style-consistency across a driving video during unsupervised video-to-video translation. Xie et al. [28] proposed an unsupervised cross-weather adaptation approach, namely OP-GAN, by integrating a self-supervised module into the CycleGAN. These two approaches are current state-of-the-art in the area, which are involved as the baseline in this study.

2.3. Mutual Information

Mutual information (MI) was first adopted to model the relationship of data from different domains by Shi and Sha [24]. They optimized the metric of MI to adapt classifiers trained on a labeled source domain to an unlabeled target domain. However, their computation of MI built upon the strong prior-knowledge, i.e., discriminative clustering, which is unable to be applied to neural networks. In more recent researches, the idea of MI has been introduced to improve the performance of GANs. InfoGAN [6] was the first work in the area, which disentangled the representative features by maximizing the mutual information between a small subset of the latent variables and the observation. Na et al. [22] proposed a mutual information loss (MILO), which optimized the same lower bound of MI, to improve the quality of synthetic images produced by the unsupervised I2I translation. Gholami et al. [13] proposed to optimize the Barber & Agakov lower bound of MI for multi-target domain adaptation. In this paper, we apply MI to deal with the distortion problem occurred in most cycle-consistency-based GANs, which is the first work in the area of cross-weather translation, to our best knowledge.

3. Mutual-GAN

Our Mutual-GAN aims to disentangle the content features from domain information for both the source and translated images, and then maximize the mutual information between the disentangled content features to preserve the image-objects. In this section, we first revisit the principle of CycleGAN and then present the modules for content feature disentanglement and mutual information maximization in details.

3.1. Revisit Cycle-consistency-based GANs

Cycle-consistency-based generative adversarial networks such as CycleGAN [31] have two paired generator-discriminator modules, which are capable of learning two
mappings, i.e., from domain A to domain B \( \{G_{AB}, D_B\} \) and vice versa \( \{G_{BA}, D_A\} \). The generators \((G_{AB}, G_{BA})\) translate images between the source and target domains, while the discriminators \((D_A, D_B)\) aim to distinguish the real and translated data. Thereby, the generators and discriminators are gradually updated during this adversarial competition.

The framework is usually supervised by two losses, i.e., adversarial loss \( (L_{adv}) \) and cycle-consistency loss \( (L_{cyc}) \). The adversarial loss encourages local realism of the translated data. Taking the translation from domain B to domain A as an example, the adversarial loss can be written as:

\[
L_{adv}(G_{BA}, D_A) = \mathbb{E}_{x_A \sim p_{x_A}} [(D_A(x_A) - 1)^2] + \mathbb{E}_{x_B \sim p_{x_B}} [(D_A(G_{BA}(x_B)))^2]
\]  \( (1) \)

where \( p_{x_A} \) and \( p_{x_B} \) denote the sample distributions of domain A and B, respectively; \( x_A \) and \( x_B \) are samples from domain A and B, respectively.

The cycle-consistency loss \( (L_{cyc}) \) tackles the requirement of paired training data. The idea behind the cycle-consistency loss is that the translated data from the target domain can be exactly converted back to the source domain, which can be expressed as:

\[
L_{cyc}(G_{BA}, G_{AB}) = \mathbb{E}_{x_A \sim p_{x_A}} [\|G_{BA}(G_{AB}(x_A)) - x_A\|_1] + \mathbb{E}_{x_B \sim p_{x_B}} [\|G_{AB}(G_{BA}(x_B)) - x_B\|_1].
\]  \( (2) \)

With these two losses, cycle-consistency-based GANs can perform image-to-image translation using unpaired training data. However, a recent study \( [31] \) found that the cycle-consistency has an intrinsic ambiguity with respect to geometric transformations. Let \( T \) be a bijective geometric transformation (e.g., translation, rotation, scaling, or even nonrigid transformation) with inverse transformation \( T^{-1} \), the following generators \( G'_{AB} \) and \( G'_{BA} \) are also cycle consistent:

\[
G'_{AB} = G_{AB}T, \quad G'_{BA} = G_{BA}T^{-1}.
\]  \( (3) \)

Consequently, due to the lack of penalty in content error between the source and translated images, an image translated from one domain to the other using cycle-consistent constraint may be geometrically distorted, as illustrated in Fig. 1. To address this problem, existing studies \( [31, 15] \) proposed to use segmentation sub-task with pixel-wise annotation, which is expensive and difficult to acquire for real applications, as an auxiliary regularization for the training of generators.

To loose the requirement of pixel-wise annotation, we propose to address the content distortion problem using a mutual information constraint. The pipeline of our approach (Mutual-GAN) is presented in Fig. 2 (a). Similar to current cycle-consistency-based GAN \( [33] \), our Mutual-GAN adopts paired generators \((G_{AB} \text{ and } G_{BA})\) and discriminators \((D_B \text{ and } D_A)\) to achieve cross-domain image translation without paired training samples. Moreover, to preserve image contents during I2I domain adaptation without the pixel-wise annotation, the proposed Mutual-GAN disentangles the content features from domain information in an unsupervised manner using pairs of content feature encoders (i.e., \( \text{Enc}_A \text{ and } \text{Enc}_B \)) and domain feature decoders (i.e., \( \text{Dec}_A \text{ and } \text{Dec}_B \)).

Specifically, the encoders (i.e., \( \text{Enc}_A \text{ and } \text{Enc}_B \)) are responsible to embed the content information of source and translated images into the same latent space \( Z \), while the decoders (i.e., \( \text{Dec}_A \text{ and } \text{Dec}_B \)) aim to transform the embedded content features to their own domains using domain-related information. Therefore, to alleviate the content distortion problem during image translation, we only need to maximize the mutual information between the content feature of source image and the translated image, which is achieved by our mutual information constraint.

### 3.2 Content Feature Disentanglement

The proposed Mutual-GAN utilizes the content feature encoders \( (\text{Enc}_A, \text{Enc}_B) \) and domain feature decoders \( (\text{Dec}_A, \text{Dec}_B) \) to disentangle the content features (i.e., image-objects) from domain information (i.e., weather and illumination). Since the mappings between domains A and
B are symmetrical, we take the content feature distillation of images from domain A as an example, as shown in Fig. 2 (b). Given an input image \( I_a \), the encoder \( Enc_A \) embeds it into a latent space, which can be formulated as:

\[
z_a = Enc_A(I_a).
\]

The embedded feature \( z_a \) contains the information of content and domain A. To disentangle them, \( z_a \) is mapped to domain A and B via \( Dec_A \) and \( Dec_B \), respectively

\[
I'_a = Dec_A(z_a), \quad I'_{ab} = Dec_B(z_a)
\]

where \( I'_a \) and \( I'_{ab} \) are the mapping results of \( z_a \) to domain A and B, respectively.

Similar to Auto-Encoder, \( I'_a \) can be seen as the reconstruction of \( I_A \). As shown in Fig. 2 apart from the X-shape dual Auto-Encoders, there is another translation path between domain A and B:

\[
I_{ab} = G_{AB}(I_a)
\]

where \( I_{ab} \) is the translated image yielded by \( G_{AB} \).

Through simultaneously minimizing the pixel-wise \( L_1 \) norm between \( I_{ab} \) and \( I'_a \), and reconstruction error between \( I_a \) and \( I_a' \), \( Dec_A \) and \( Dec_B \) are encouraged to recover domain-related information from the latent space (in short, the encoders remove domain information and the decoders recover it), which enables them to map the \( z_a \) to two different domains. Therefore, the information contained in \( z_a \) is highly related to the image-objects without domain bias. These content feature distillation loss \( L_{dis} \), combining aforementioned two terms, can be formulated as:

\[
L_{dis} = ||I_{ab} - I'_a||_1 + ||I_a - I_a'||_1.
\]

### 3.3. Mutual Information Maximization

Using the content feature encoder and domain feature decoders, the content feature of source image \( I_a \) can be embedded to \( z_a \). To preserve image-objects during I2I translation, the translated image \( I_{ab} \) is enforced to contain similar information to the embedded content feature \( (z_a) \), which can be achieved by maximizing the mutual information \( I(z_a) \) between \( I_{ab} \) and \( z_a \). Therefore, we can formulate the mutual information loss \( L_{MI} \) as:

\[
L_{MI} = I(z_a; I_{ab}) = H(z_a) - H(z_a|I_{ab})
= E_{x~I_{ab}}[log P(z_a|x)] + H(z_a).
\]

However, according to [6], the posterior probability \( P(z_a|x) \) is difficult to be directly estimated. To this end, we utilize an auxiliary distribution \( Q(.) \), which is easily accessed, to approximate the posterior probability:

\[
I(z_a; I_{ab}) = E_{x~I_{ab}}[log P(z'_a|x)] + H(z_a)
= E_{x~I_{ab}}[KL(P(.|x)||Q(.|x)) + E_{z'_a~P(z_a|x)}log Q(z'_a|x)] + H(z_a).
\]

where \( KL \) is the KL-divergence.

Since \( KL(P(.|x)||Q(.|x)) \geq 0 \), we can obtain the lower bounded mutual information based on variational information maximization [11]:

\[
I(z_a; I_{ab}) \geq E_{x~I_{ab}}[log Q(z'_a|x)] + H(z_a).
\]

Hence, the proposed Mutual-GAN can maximize the \( I(z_a; I_{ab}) \) via optimizing the lower bound—the lower bound becomes tight as the auxiliary distribution \( Q \) approaches the true posterior distribution: \( E_{x}[KL(P(.|x)||Q(.|x))] \rightarrow 0 \).

In practice, \( E_{x~I_{ab}}[log Q(z'_a|x)] \) can be parametrized using \( Enc_B; z'_a \rightarrow z_{ab} \), as illustrated in Fig. 2 (c). Therefore, the maximization of \( I(z_a; I_{ab}) \) can be achieved by aligning the distributions of \( P(z_a|x) \) and \( Q(z_{ab}|x) \) with \( E_{x~I_{ab}} \). In our experiments, the \( L_1 \) norm is adopted for the distribution alignment:

\[
L_{MI} = ||z_a - z_{ab}||_1.
\]

### 3.4. Objective

With the previously defined feature distillation loss \( L_{dis} \) and mutual information constraint \( L_{MI} \), the full objective \( L \) for the proposed Mutual-GAN is summarized as:

\[
L = L_{adv}(G_{BA}, D_A) + L_{adv}(G_{AB}, D_B)
+ \alpha L_{cyc}(G_{BA}, G_{BA})
+ L_{dis}(G_{BA}, Enc_A, Dec_A, Dec_B)
+ L_{dis}(G_{BA}, Enc_B, Dec_B, Dec_A)
+ L_{MI}(G_{BA}, Enc_A, Enc_B) + L_{MI}(G_{BA}, Enc_A, Enc_B)
\]

where \( L_{adv} \) and \( L_{cyc} \) are adversarial and cycle-consistency losses respectively, the same as that proposed in [33].

The optimization of \( L_{dis} \) and \( L_{MI} \) is performed in the same manner of \( L_{adv} \)—fixing \( Enc_A, Enc_B, Dec_A, Dec_B, \) and \( D_A, D_B \) to optimize \( G_BA, G_{AB} \) first, and then optimize \( Enc_A, Enc_B, Dec_A, Dec_B, \) and \( D_A, D_B \) respectively, with \( G_BA, G_{AB} \) fixed. Therefore, similar to discriminators, our content encoders and domain decoders can directly pass the knowledge of image-objects to the generators, which helps them to improve the quality of translated results in terms of object preservation.

### 3.5. Comparison with CycleGAN and InfoGAN

We notice that our Mutual-GAN is highly related to CycleGAN [33] and InfoGAN [6]: therefore, it is worthwhile to emphasize the differences between the proposed model and the existing frameworks.

Compared to CycleGAN, our Mutual-GAN consists of two novel components: content feature disentanglement and mutual information maximization modules, which alleviate the problem of image-object distortion via maximizing the mutual information between content features \( z_a \) and
In comparison with InfoGAN, we propose a new paradigm to measure the mutual information (MI) between a feature ($z_b$) and an image ($I_{ab}$), since the proposed Mutual-GAN involves more network components for the MI calculation ($Enc_A, G_{AB}, Enc_B$ vs. generator-only in InfoGAN).

4. Experiments

As aforementioned, semantic segmentation networks trained with standard scenario images (domain A), i.e., daytime scenes with favorable illuminations, suffer from performance degradation while testing on real images under changeable weathers (domain B), e.g., cloudy and nighttime. Hence, our goal is to narrow down the gap between images under different weathers not only in terms of visual perception i.e., plausible translated results, but also the representation in feature space, i.e., improvement of the robustness of models. We present the image-to-image translation results to evaluate the former factor. For the latter one, we evaluate the Mutual-GAN in the similar transfer learning scenario. To conduct experiments on cloudy-to-sunny translation to evaluate our Mutual-GAN.

4.1. Datasets

Experiments are conducted on two publicly available datasets to demonstrate the effectiveness of our Mutual-GAN. The datasets are:

CamVid. The dataset contains driving videos under different weathers, e.g., cloudy and sunny. The task translating cloudy videos to the sunny domain is very challenging, as the cloudy videos are often very dark, which lose much detailed information. We conduct experiments on cloudy-to-sunny translation to evaluate our Mutual-GAN.

SYNTHIA. The dataset consists of photo-realistic frames rendered from a virtual city. The night-to-day translation is a more difficult task than the cloudy-to-sunny, since the night domain suffers from severe loss of context information. We examine how the proposed Mutual-GAN performs on the night-to-day translation task using two sub-sequences (i.e., summer (daytime) and night (nighttime)) from the Highway, which is a subset of SYNTHIA.

4.2. Experiment Settings

Evaluation criterion. The mean of class-wise intersection over union (mIoU) is used to evaluate the improvement achieved by our Mutual-GAN on the semantic segmentation tasks for CamVid and SYNTHIA datasets.

Baselines overview. Several unpaired image-to-image domain adaptation frameworks, including CycleGAN, UNIT, DRIT and OP-GAN are taken as baselines for the performance evaluation. The state-of-the-art frameworks (i.e., VideoGAN and OP-GAN) and the direct transfer approach, which directly takes the source domain data for testing without any adaptation, are also involved for comparison. Note that the recently proposed GAN for image-based domain adaptation, i.e., AugGAN, is not involved for comparison, due to the strong prior-knowledge (pixel-wise annotation) used in the approach.

The benchmarking algorithms adopted in this study is the same to the previous studies. Some of the existing unsupervised domain adaptation (UDA) methods are not involved for comparison, since they are feature-alignment (FA) based, while our Mutual-GAN is an image-to-image-based domain adaptation method. The difference between these two types of UDA methods is I2I-based methods are task-agnostic, compared to the FA-based domain adaptation. In other words, the FA-based methods require specific-annotations (e.g., pixel-wise annotation) of source domain to perform adaptive domain adaptation, which is not a necessary requirement for I2I-based approaches. To this end, it is difficult to fairly compare the performances between these two types of UDA approaches.

Training details. The proposed Mutual-GAN is implemented using PyTorch. The generator, discriminator, and Siamese network are iteratively trained for 200 epochs with the Adam solver. The split of training, validation and test sets for semantic segmentation on both CamVid and SYNTHIA datasets is consistent to . The baselines involved in this study adopt the same training protocol.

4.3. Visualization of Translation Results

The translation results for the two tasks generated by different image-to-image translation frameworks are presented in Fig. which illustrate the main problem of existing approaches (UNIT, DRIT and CycleGAN) — image content corruptions. Due to the lack of penalty in content error between the source and translated images, the existing translation frameworks intend to overly edit the image content such as changing the shape and colors of image-objects, referring to the road and building in the CamVid and SYNTHIA translated images. This drawback restricts

1The transfer learning scenario is more practical for realistic applications, as the well-trained deep learning networks can be directly applied to unlabeled data from another domain. In contrast, the data augmentation scenario still needs the pixel-wise annotation to re-train the networks.  
2http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/  
3http://synthia-dataset.net/  
4Since VideoGAN is only evaluated on CamVid, we reimplement the framework and report the performance of our reproduction. Instead, OP-GAN has been evaluated on both datasets adopted in this study; hence, we involve the accuracy reported in for comparison.
Figure 3. Comparison of cross-weather adaptation results yielded by different approaches. The image adaptation performance is evaluated on two tasks (cloudy-to-sunny and nighttime-to-daytime). The original images, translated results produced by UNIT [20], DRIT [18], CycleGAN [33], VideoGAN [5] and our Mutual-GAN are presented. Since we do not reproduce the OP-GAN [28], its translation results are excluded from the figure.

Table 1. Semantic segmentation IoU (%) of the cloudy images from CamVid with different image-to-image translation frameworks.

| Sunny (validation) | Bicyclist | Building | Car | Pole | Fence | Pedestrian | Road | Sidewalk | Sign | Sky | Tree | mIoU |
|--------------------|-----------|----------|-----|------|-------|------------|------|----------|------|-----|------|------|
| Direct transfer    | 85.30     | 85.64    | 90.40 | 20.60 | 76.00 | 61.82      | 94.46 | 89.48    | 12.12 | 94.20 | 91.83 | 71.19 |
| UNIT [20]          | 0.67      | 49.28    | 6.99 | 6.10 | 1.54  | 0.79       | 45.05 | 15.16    | 0.00  | 62.61 | 39.34 | 19.94 |
| DRIT [18]          | 0.34      | 40.00    | 0.31 | 0.33 | 0.83  | 0.23       | 48.27 | 26.81    | 0.00  | 64.85 | 23.73 | 18.20 |
| CycleGAN [33]      | 6.16      | 56.64    | 10.76 | 8.61 | 0.01  | 4.06       | 50.49 | 30.21    | 8.67  | 75.97 | 45.62 | 26.26 |
| VideoGAN [5]       | 33.57     | 73.22    | 43.08 | 22.79 | 12.39 | 34.70      | 68.83 | 48.65    | 22.09 | 76.38 | 73.10 | 46.37 |
| OP-GAN [28]        | 51.28     | 73.10    | 74.19 | 25.84 | 12.42 | 42.75      | 70.48 | 51.74    | 14.71 | 81.09 | 72.40 | 51.40 |
| Mutual-GAN (Ours)  | 48.15     | 69.51    | 60.72 | 25.95 | 15.04 | 42.75      | 81.44 | 69.49    | 10.84 | 91.78 | 70.63 | 51.93 |
| Ours w/o LMI       | 31.51     | 67.12    | 14.17 | 18.85 | 6.73  | 23.54      | 73.49 | 59.33    | 19.05 | 87.05 | 73.33 | 41.82 |

4.4. Comparison with State-of-the-art

In this section, we evaluate the performance of our Mutual-GAN on the two I2I domain adaptation tasks and compare with the state-of-the-art approaches, e.g., VideoGAN [5] and OP-GAN [28].

Cloudy-to-sunny adaptation on CamVid. The CamVid dataset contains four sunny videos (577 frames in total) and one cloudy video (124 frames). Each frame of the videos is manually annotated, which associates each pixel with one of the 32 semantic classes. Based on the widely-accepted protocol [34], we focus on 11 classes including bicyclist, building, car, pole, fence, pedestrian, road, sidewalk, sign, sky, and tree. To evaluate the translation performance yielded by our Mutual-GAN, a semantic segmentation network (PSPNet) is trained with the sunny frames and tested on the original cloudy frames and the translated ones. In the experiment, the sunny frames are separated to training (three videos) and validation (one video) sets. The evaluation results are shown in Table 1.

Due to the loss of detailed information, it can be observed from Table 1 that the performance of PSPNet trained with sunny images dramatically drops to 39.48% while tested on the original cloudy images. As the existing I2I...
domain adaptation approaches may encounter the content distortion problem and corrupt the image-objects, the semantic segmentation mIoU of PSPNet further degrades to 26.26%, 19.94% and 18.20% using the CycleGAN, UNIT and DRIT, respectively. The video-to-video domain adaptation approach (VideoGAN) marginally improves the mIoU to 46.37%, compared to the direct transfer. In contrast, the proposed Mutual-GAN significantly boosts the accuracy of most classes, especially for the sky (i.e., +23.18%). The mIoU of our Mutual-GAN (i.e., 51.93%) is slightly higher than the state-of-the-art OP-GAN (51.40%), which demonstrates its effectiveness on image-object preservation.

**Nighttime-to-daytime adaptation on SYNTHIA.** We adopt two sub-sequences (summer (daytime) and night (nighttime)) from SYNTHIA to perform nighttime-to-daytime adaptation. The daytime and nighttime sequences contain 603 and 461 frames, respectively. SYNTHIA dataset provides pixel-wise semantic annotations for each frame, which can be categorized to nine semantic classes, as listed in Table 2. The partition of training, validation and test sets complies with the same protocol to that of the CamVid dataset—the daytime images are separated to training and validation sets according to the ratio of 70:30, while all the nighttime images are used for test. The fully convolutional network (PSPNet) is also adopted in this experiment to perform semantic segmentation.

The semantic segmentation mIoUs yielded by different testing strategies are shown in Table 2. Similar to the cloudy images, the PSPNet trained with daytime images fails to properly process the nighttime images—an mIoU of 47.36%, due to the loss of information. The nighttime-to-daytime translation on SYNTHIA is a more challenging task compared to the cloudy-to-sunny translation on CamVid, since large areas of night images are dark, where the image-objects (e.g., cars) are difficult to recognize. Refer to Fig. 3 most of existing image-to-image translation frameworks often create contents to fill the extremely dark areas, which consequently corrupts the original image-objects, e.g., cars and buildings. Due to these distortions, the images translated by CycleGAN and DRIT further decrease the mIoU of PSPNet to 22.51% and 35.01%, respectively. The OP-GAN, VideoGAN and UNIT yield improvements to mIoU, i.e., +1.61%, +7.31% and +13.85%, respectively. Since our Mutual-GAN excellently prevents image-object corruptions during nighttime-to-daytime translation (refer to Fig. 3), it achieves the best mIoU (67.27%) for nighttime images, which is comparable to the validation result (70.06%) tested on daytime images.

**Visualization of segmentation results.** The segmentation results yielded by PSPNet with different I2I adaptation approaches are presented in Fig. 4. It can be observed that our Mutual-GAN can significantly improve the quality of semantic segmentation results of images under unfavorable weathers, which illustrates its potential for real applications.

| Sky | Building | Road | Fence | Veg. | Pole | Car | Sign | 1.-m. mIoU |
|-----|----------|------|-------|------|------|-----|------|-----------|
| Day (validation) | 94.20    | 80.83 | 97.76 | 89.34 | 67.61 | 40.42 | 66.98 | 2.41       |
| Night (test)    |          |       |       |       |      |      |      |           |
| Direct transfer | 25.67    | 58.49 | 89.16 | 64.44 | **57.53** | 39.32 | 10.20 | 23.13     |
| UNIT [20]       | 91.93    | **77.02** | 90.63 | 29.12 | 55.84 | 39.84 | **74.34** | 28.16     |
| DRIT [18]       | 77.25    | 39.12 | 72.28 | 4.33  | 43.47 | 25.80 | 36.57 | 6.22      |
| CycleGAN [33]   | 63.72    | 22.21 | 59.05 | 0.89  | 24.63 | 15.75 | 2.36  | 0.00      |
| VideoGAN [5]    | 91.59    | 43.29 | 88.78 | 49.56 | 39.59 | 33.84 | 45.24 | 15.11     |
| OP-GAN [23]     | 21.90    | 66.22 | 86.78 | 7.7   | 54.86 | 39.11 | 85.09 | 31.40     |
| Mutual-GAN (Ours) | **94.67** | 70.54 | **92.23** | 69.73 | 54.64 | **41.22** | 69.05 | 30.29     |
| Ours w/o \(L_M\) | 89.53    | 48.91 | 91.11 | 41.28 | 56.29 | 34.32 | 52.50 | 10.88     |

Table 2. Semantic segmentation IoU (%) of the night images from SYNTHIA with different image-to-image translation frameworks. (Veg. – Vegetation, 1.-m. – lane-marking)
Figure 4. Comparison of segmentation results yielded by PSPNet using different I2I adaptation approaches. The original images, segmentation results using UNIT [20], DRIT [18], CycleGAN [33], VideoGAN [5] and our Mutual-GAN are presented. Since we do not reproduce the OP-GAN [28], its segmentation results are excluded from the figure. (G. T.–Ground truth, D. T.–Direct transfer)

Figure 5. Attention maps generated by $\text{Enc}_B$ and $\text{Enc}_A$ for cloudy and translated images, respectively.

Mutual information constraint. To validate the effectiveness of our mutual information constraint, we conduct an ablation study on the two datasets. The evaluation results are presented in Table 1 and Table 2. The mIoU of PSPNet remarkably degrades while testing on the images translated by Mutual-GAN without $L_{MI}$—mIoU degrades to 41.82% and 56.57% on CamVid and SYNTHIA, respectively. The experimental results demonstrate that our mutual information constraint effectively enforces the Mutual-GAN to preserve the image-objects during cross-weather adaptation.

5. Conclusion

In this paper, we propose a novel generative adversarial network (namely Mutual-GAN) to alleviate the accuracy decline when daytime-trained neural network is applied to videos captured under adverse weather conditions. The proposed Mutual-GAN adopts mutual information constraint to preserve image-objects during cross-weather adaptation, which is an unsolved problem for most unsupervised image-to-image translation approaches (e.g., CycleGAN). The proposed Mutual-GAN is evaluated on two publicly available driving video datasets (i.e., CamVid and SYNTHIA). The experimental results demonstrate that our Mutual-GAN can yield visually plausible translated images and significantly improve the semantic segmentation accuracy of daytime-trained deep learning network while processing videos under challenging weathers.

Limitation & Future work. We notice the limitation of this study: the weathers are translated only under two scenarios, i.e., cloudy vs. sunny and nighttime vs. daytime, and plan to resolve it in our future work.
References

[1] David Barber and Felix V. Agakov. The im algorithm: A variational approach to information maximization. In Conference on Neural Information Processing Systems, 2003.

[2] Javad Zolfagharí Bengar, Abel Gonzalez-Garcia, Gabriel Villalonga, Bogdan Raducanu, Hamed A. Aghdam, Mikhail Mozerev, Antonio M. Lopez, and Joost van de Weijer. Temporal coherence for active learning in videos. arXiv preprint arXiv:1908.11757, 2019.

[3] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In European Conference on Computer Vision, 2008.

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[5] Jiawei Chen, Yuexiang Li, Kai Ma, and Yefeng Zheng. Generative adversarial networks for video-to-video domain adaptation. In AAAI Conference on Artificial Intelligence, 2020.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Conference on Neural Information Processing Systems, 2016.

[7] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. CartoonGAN: Generative adversarial networks for photo cartoonization. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[8] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometricaly guided input-output adaptation approach. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[9] Yuhua Chen, Wen Li, and Luc Van Gool. ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[10] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation. In International Conference on Computer Vision, 2019.

[11] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In International Conference on Intelligent Transportation Systems, 2018.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[13] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Boumalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. IEEE Transactions on Image Processing, 29:3993–4002, 2020.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Conference on Neural Information Processing Systems, 2014.

[15] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. AugGAN: Cross domain adaptation with GAN-based data augmentation. In European Conference on Computer Vision, 2018.

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In International Conference on Machine Learning, 2017.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In European Conference on Computer Vision, 2018.

[19] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[20] Ming-Yu Liu, Chang Wen Chen, and Tao Mei. TransGaAN: Instance-level image translation via deep attention generative adversarial networks. In Conference on Neural Information Processing Systems, 2017.

[21] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[22] Sanghyeon Na, Seungjoo Yoo, and Jaegul Choo. MISO: Mutual information loss with stochastic style representations for multimodal image-to-image translation. arXiv preprint arXiv:1902.03938, 2019.

[23] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[24] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In International Conference on Machine Learning, 2012.

[25] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[26] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. Stellar: Real-time cross-domain semantic segmentation through day-night image conversion. In Artificial Intelligence and Machine Learning in Defense Applications, volume 1169, 2019.

[27] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[28] Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. Self-supervised CycleGAN for
object-preserving image-to-image domain adaptation. In *European Conference on Computer Vision*, 2020.

[29] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dual-GAN: Unsupervised dual learning for image-to-image translation. In *International Conference on Computer Vision*, 2017.

[30] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Conference on Neural Information Processing Systems*, 2019.

[31] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and-shape-consistency generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, 2017.

[34] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[35] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE International Conference on Computer Vision*, 2019.