This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

Panoptic-Level Image-to-Image Translation for Object Recognition and Visual Odometry Enhancement

Liyun Zhang, *Graduate Student Member, IEEE*, Photchara Ratsamee, *Member, IEEE*, Zhaojie Luo, *Member, IEEE*, Yuki Uranishi, *Member, IEEE*, Manabu Higashida, Haruo Takemura, *Member, IEEE*

***Abstract*—Image-to-image translation methods have progressed from only considering the image-level information to integrating the global- and instance-level information. However, only the foreground instances are refined, and the background semantics are taken as an entire feature, which causes a substantial loss of the semantic information in the translation. Additionally, the insufficient quality of the translated semantic regions also leads to an unsatisfactory performance of the object recognition or visual odometry tasks in which the translated images/videos are further used. In this paper, we propose a novel generative adversarial network for panoptic-level image-to-image translation (PanopticGAN). The proposed method has three advantages: (1) the extracted panoptic perception (i.e., the foreground instances and background semantic regions) as content codes are aligned with the sampled panoptic style codes, which considers the panoptic-level information to avoid the semantic information loss, and the latent space of each object has a rich fusion of content and style codes to generate the higher-fidelity results; (2) a feature masking module is proposed to extract the representations within each object contour by masks for sharpening the object boundaries; (3) the improved fidelity of the translated semantic regions further contributes to enhancing the performance of the object recognition or visual odometry tasks that the translated images/videos are used in. In this paper, we also annotate a compact panoptic segmentation dataset for the thermal-to-color translation task. Extensive experiments are conducted to demonstrate the effectiveness of our PanopticGAN over the latest methods.**

***Index Terms*—Image-to-image translation, panoptic-level, feature masking, generative adversarial networks, image/video enhancement.**

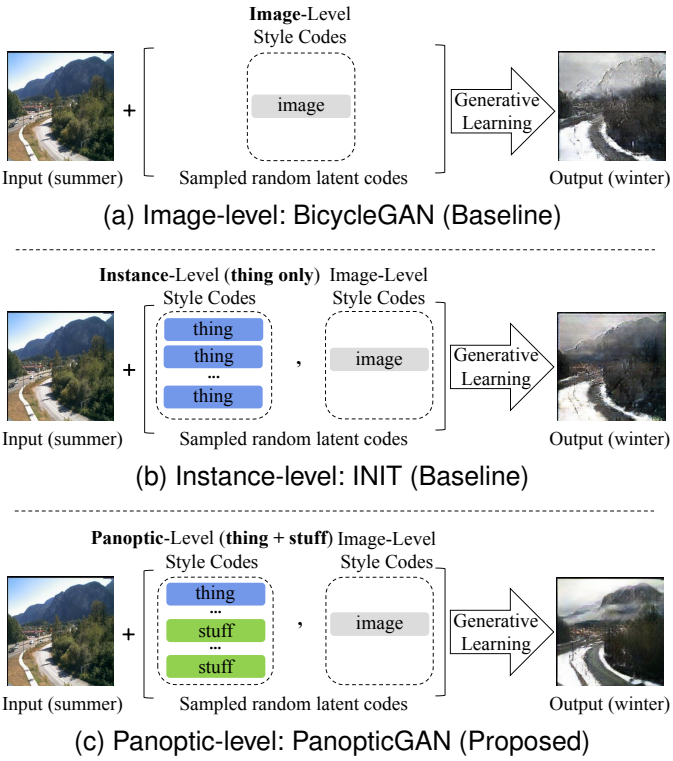


Fig. 1. Pipeline comparisons of image-level baselines [1], instance-level baselines [2], and our proposed panoptic-level approach. (a) only uses image-level style codes randomly sampled from the style space of the target domain for image translation; (b) combines instance-level style codes (target objects are the only countable foreground instances 'thing', e.g., car or traffic light) and image-level style codes; Our proposed approach (c) combines panoptic-level style codes (target objects are both 'thing' and uncountable background segments 'stuff', e.g., tree or road) and image-level style codes to avoid losing too much information in the translation.

I. Introduction

Manuscript received August 13, 2022; revised February 22, 2023 and April 28, 2023; accepted June 10, 2023. This work has been partly supported by the KAKENHI Fund for the Promotion of Joint International Research (fostering joint international research (B) No. 20KK0086) and the Mohamed Bin Zayed International Robotics Challenge (MBZIRC) Grant. (Corresponding author: Liyun Zhang, Photchara Ratsamee and Zhaojie Luo.)

Liyun Zhang is with the Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan (e-mail: liyun.zhang@lab.ime.cmc.osaka-u.ac.jp).

Photchara Ratsamee is with the Faculty of Robotics and Design, Department of System Design, Osaka Institute of Technology, Osaka 530-8568, Japan (e-mail: photchara@ime.cmc.osaka-u.ac.jp).

Zhaojie Luo is with the Department of Information and Communications Technology, Osaka University, Osaka 565-0871, Japan (e-mail: luo@sanken.osaka-u.ac.jp).

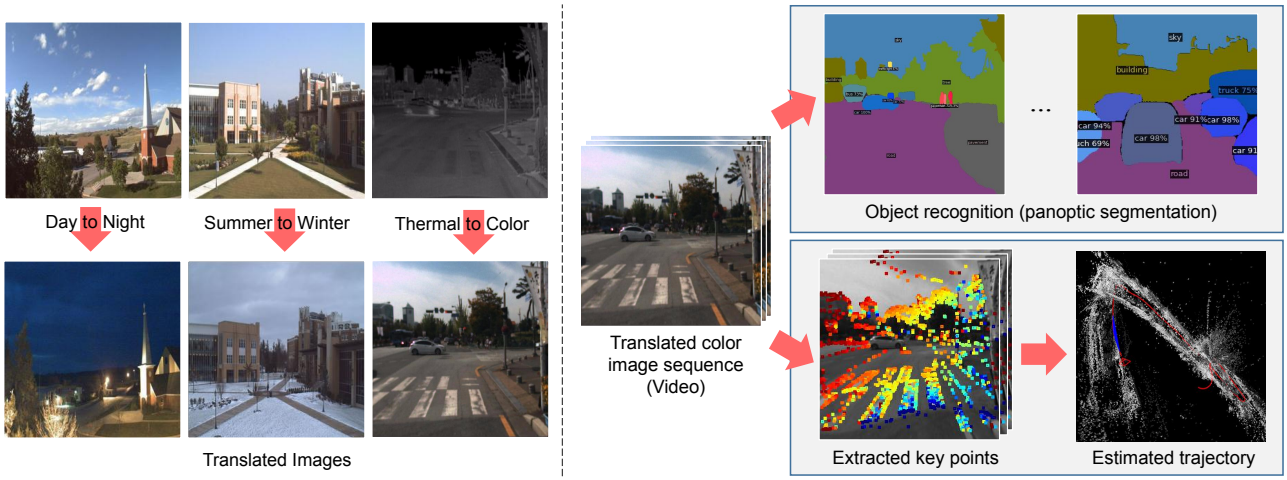
Yuki Uranishi, Manabu Higashida, and Haruo Takemura are with the Infomedia Education Division, Cybermedia Center, Osaka University, Osaka 565-0871, Japan (e-mail: yuki.uranishi.cmc@osaka-u.ac.jp; manabu@cmc.osaka-u.ac.jp; takemura@cmc.osaka-u.ac.jp).

IMAGE -to-image (I2I) translation is a challenging problem in the computer vision field. Translated images need to retain the content information of the input domain image and obtain the style of the target domain [3], [4]. Several tasks can be considered I2I translation problems, e.g., superresolution [5], [6], neural style transfer [7], and colorization [8], [9]. Utilizing the enhanced images/videos through I2I translation to further improve the task performance has become a prominent area of research in recent years, e.g., Fu et al. [10] utilized a night-to-day image translation to enhance the object detection performance at nighttime; Zhu et al. [11]

Copyright © 20xx IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547



(a) Panoptic-level image-to-image translation (b) Enhancement for object recognition and visual odometry

Fig. 2. Our motivation consists of two parts: (a) is our proposed method, which is used for general I2I translation tasks, e.g., summer-to-winter, day-to-night and thermal-to-color, etc.; (b) indicates the mission of our paper. We expect that using our translated image sequences can enhance the performance of the object recognition tasks and improve the extracted key points for better a performance (more accurate estimated trajectory) in the visual odometry tasks. The white and semidense dots represent the 3D reconstructed point clouds, and the red curve represents the estimated trajectory.

proposed a night-to-day image translation for the performance enhancement of background modeling; ForkGAN [12] presented a task-agnostic image translation to boost multiple vision tasks; MFGAN [13] transferred night videos to day videos to improve the resolution for impressive gains in the visual odometry and simultaneous localization and mapping (SLAM) [14] tasks. However, these methods only consider the image-level information in the I2I translation without refining the object semantics, or the advanced instance-level methods proposed later also only refined the foreground instance objects without considering the background semantic regions. Due to a substantial loss of the semantic information in the translation, it causes the insufficient quality of the translated objects and further leads to a limited enhancement of the task performance. Therefore, how to avoid the semantic information loss to translate the higher-fidelity images for improving the task performance is a fundamental problem and an important challenge.

I2I translation has evolved from image-level methods [15], [16] to instance-level methods [2], resulting in significant improvements in the fidelity of translated images. As illustrated in Fig. 1 (a), the image-level baseline extracts the image representations as content codes to combine with the image-level style codes, which are randomly sampled from the style space of the target domain for I2I translation. The instance-level baseline in Fig. 1 (b) uses a pretrained instance segmentation network [17] to extract the instance perception from the input image, which contains object bounding boxes, categories, and masks. It uses Region of Interest Align (RoIAlign) [17] to extract the instance representations, which combines the image representations as content codes. The sampled instance-level style codes and image-level style codes are aligned with the corresponding content codes for an instance-level I2I translation. Compared with image-level I2I translations, instance-level I2I translations can refine the foreground object instance representation precisely, however, the background semantic regions are not fully refined. In contrast, panoptic-level I2I translations can thoroughly refine the foreground object instances 'thing' and background semantic regions 'stuff' [18] in the translation for preventing the loss of too much information. From a theoretical perspective, the panoptic-level method captures an adequate amount of information to achieve higher fidelity in both the foreground and background.

In this paper, our proposed PanopticGAN uses a pretrained panoptic segmentation [19] network to extract the panoptic perceptions. In Fig. 1 (c), 'thing' (foreground object instances) and 'stuff' (background semantic regions) representations are extracted as object content codes to avoid the semantic information loss, which are combined with the image-level representations as content codes. The sampled panoptic-level style codes and image-level style codes are aligned with the corresponding content codes to generate the higher-fidelity images during the translation process. We also propose a novel feature masking module that leverages object contour masks to extract object-specific representations and sharpen object boundaries. Therefore, our proposed method improves the fidelity of the translated semantic regions and enhances the entire image quality, which contributes to further enhancing the performances of the object recognition and visual odometry by using the translated images/videos.

Fig. 2 illustrates the mission of this paper. In Fig. 2 (a), our proposed method can be used for general I2I translation tasks, e.g., summer-to-winter, day-to-night and thermal-to-color, etc., which can obtain higher image quality (realism, sharpness and diversity) than the baseline approaches. In Fig. 2 (b), the original (thermal) subsequences are translated to a target (color) subsequences for a performance enhancement in object recognition and visual odometry tasks. The top right shows the enhanced object recognition results on the panoptic segmentation [19], which recognized the objects more

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

exhaustively and accurately. In the bottom right, the Direct Sparse Odometry (DSO) [20] method is run on the translated subsequences (video) to extract the enhanced key points with depth maps, which denote more robust gradient points for better 3D reconstructions (point clouds) and estimated scene movement (trajectories). Our main research contributions are fourfold:

- A pioneering panoptic-aware I2I translation network is proposed to refine both the foreground instances and background semantic regions, which avoids a substantial loss of the semantic information, and the latent space of each object has a rich fusion of content and style codes to generate higher-fidelity images.
- A feature masking module is proposed to extract the representations within each object contour by masks; it removes the redundant background information that exists in the object features cropped by the bounding box for sharpening object boundaries and translating higher fidelity objects.
- The proposed panoptic-level model improves the fidelity of the translated semantic regions and enhances the entire image quality, by using the translated images/videos for the object recognition and visual odometry tasks can enhance the performances compared to the original images/videos.
- We also annotated a compact panoptic segmentation dataset on a partial KAIST-MS dataset [21] for the thermal-to-color translation task. Extensive experiments are conducted on the image quality comparisons, the performance enhancement of the object recognition and visual odometry tasks, including ablation studies, to demonstrate the superiority of our proposed approach.

Our preliminary work (Zhang et al. [22]) has been published in a previous conference WACV 2023. Compared to the previous conference paper, our major new contributions are that we extend the image synthesis via the proposed image translation algorithm to continuous sequence frames (video) and use the results to enhance the monocular visual odometry performance, which makes this work a complete benchmark. We propose a novel panoptic-level image-to-image translation approach to boost the previous state-of-the-art methods, and the synthesized higher fidelity images/videos can be used in object recognition and visual odometry to achieve an enhanced and fairly competitive performance, respectively. We added comparisons of the extracted key points and estimated trajectories from the translated visual odometry image sequences (video). Moreover, we also conduct additional experiments by adding new baselines and competing models improved by the panoptic information for evaluation comparisons.

The remainder of this paper is organized as follows. The related works are reviewed in Section II. The proposed module and the presented translation framework are described in Section III. Section IV presents and discusses experiments, results and corresponding analyses including ablation studies and model efficiencies. Finally, conclusions and future works are provided in Section V.

II. Related Works

A. Image-to-Image Translation

Image-to-image (I2I) translation models transform the input domain image to the target domain [3], changing the style but keeping the scene content unchanged. Pix2Pix [15] can achieve paired mapping learning, however, it needs the paired datasets and generates a single-modal output. BicycleGAN [1] can generate multimode and more diverse results by encouraging a bijective mapping between the latent and output spaces. CycleGAN [23] can achieve unpaired dataset training by using cycle consistency loss. The disentangled representation models [3], [24] utilize a combination of the content from the input domain and the style from the target domain for unsupervised learning. TICCGAN [25] adds perceptual loss [7] and total variation (TV) loss [26] to optimize networks. Pix2PixHD [16] can translate high-resolution images via a multiscale discriminator and coarse-to-fine generator. PGGAN [27] utilizes a progressive growth strategy to synthesize images from low to high resolution. AGGAN [28] and U-GAT-IT [29] can extract attention regions as structure guidance to localize important content for high-quality results. TSIT [30] uses a two-stream generative model with a feature transformation in a coarse-to-fine fashion for capturing and fusing the multiscale semantic structure information and style representation. FDIT [31] proposes a novel frequency domain image translation framework, exploiting the frequency information to enhance the image generation process. Pang et al. [32] proposed a unified GAN network that jointly estimates transmission maps, atmospheric light, and haze-free images for image dehazing, which can be regarded as a type of image-level translation from hazy to haze-free. The pSp [33] uses a novel encoder to directly map a real image into the W+ latent space with no optimization needed, and styles are extracted in a hierarchical fashion and fed into the corresponding inputs of a fixed StyleGAN [34] generator. However, the above image-level I2I translation methods only consider the image-level information not refine the object-level features in the translation.

B. Instance-Level Image-to-Image Translation

The instance-level I2I translation was derived from the development of object-driven image generation research, e.g., Scene-Generation [35] and sg2im [36] can synthesize images from object scenes, and Layout2Im [37], OC-GAN [38], and LostGAN [39] can utilize scene layouts to achieve object-level image synthesis. These object-driven methods transform nonimage containers (e.g., scene layouts with bounding boxes) with object style codes to the target image, however, adopting bounding boxes to define objects makes it difficult to generate objects with sharp boundaries. The instance-level I2I translation methods use object perception (bounding boxes or masks), which is effective for generating sharp object boundaries. Instagan [40] incorporates a set of instance attributes for instance-aware I2I translation. DA-GAN [41] uses a deep attention encoder to enable the instance-level correspondences to be discovered. SCGAN [8] and SalG-GAN [42] utilize saliency-based guidance for image translation, which prioritizes salient regions and is considered an instance-level method

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

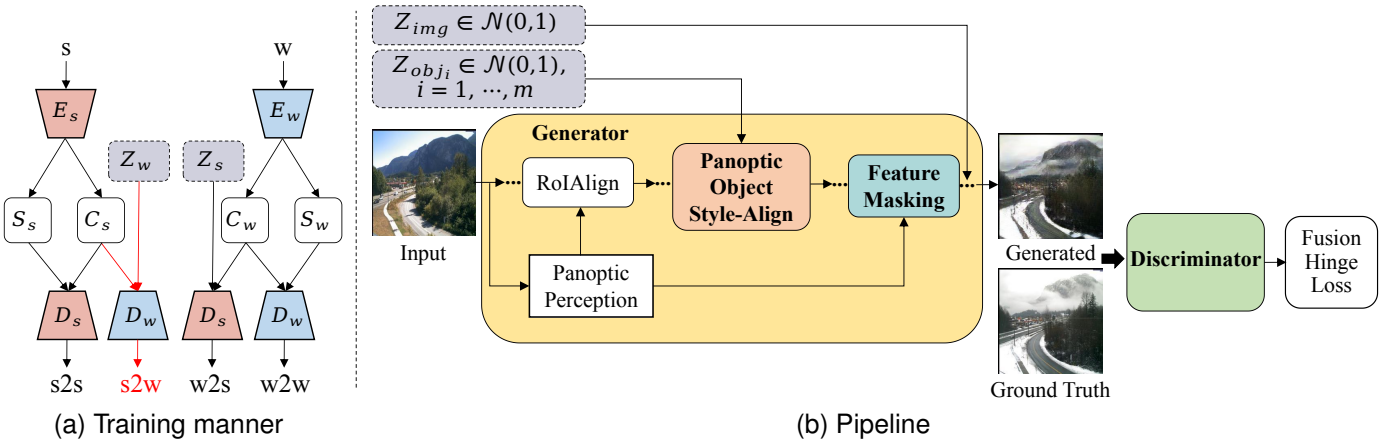


Fig. 3. Overview. In (a), the summer s and winter w images are self-encoded ($s2s$ and $w2w$) and cross-encoded ($s2w$ and $w2s$) simultaneously for image translation, where the style codes Z_w and Z_s are randomly sampled from the normal distribution. The red arrows in (a) correspond to the process of (b). In (a), a panoptic perception is extracted from the pretrained panoptic segmentation model to support the basic RoIAlign module and proposed modules (panoptic object style-align and feature masking). The sampled image-level Z_{img} and panoptic-level Z_{obj} style codes are combined for target image generation.

due to the higher saliency values of foreground objects. Shen et al. [2], Su et al. [43] and Chen et al. [44] combined an instance-level feature with an image-level feature for higher quality instance-level I2I translation. InstaFormer [45] presents a novel Transformer-based network instance-aware image-to-image translation network to effectively integrate the global-level and instance-level information. However, these instance-level I2I translation methods only refine the foreground instance objects without considering the background semantic regions, resulting in a significant loss of the background semantic information.

C. Panoptic-Level Image-to-Image Translation

To the best of our knowledge, the panoptic-level I2I translation problem has not yet been investigated. From a theoretical perspective, instance-level I2I translations only consider foreground instances as objects for learning, and have certain disadvantages compared with panoptic-level I2I translations, which regards both the foreground 'thing' and background 'stuff' as objects to avoid losing too much information in the translation. Lin et al. [46] extracted image regions for the discriminator to improve the performance of the GANs. Huang et al. [47] provided a solution to semantically control output based on references. Dundar et al. [48] proved that panoptic-level perception makes the generated images have a higher fidelity and shows the tiny objects in more detail. Semantic segmentation [49]–[52] can obtain a semantic perception of an image, but it cannot identify instances. Instance segmentation [53]–[56] can obtain the object instance perception of an image, but it cannot segment the uncountable background regions. Panoptic segmentation [18], [57] combines semantic segmentation and instance segmentation to define the uncountable background regions (e.g., sky and road) as 'stuff' and the countable foreground instances (e.g., person and car) as 'thing' for a thorough image perception. Therefore, we use a pretrained panoptic segmentation [19] network to extract panoptic perception (covering 'thing' and 'stuff') to combine the sampled panoptic-level style codes and image-level style codes with the corresponding content codes for higher-fidelity panoptic-aware I2I translations. Compared to the image-level and instance-level methods, our proposed panoptic-level method refines the object semantics on both the foreground and background to avoid the semantic information loss, and improves the fidelity of the translated objects to futher enhance the task performance.

III. The Proposed Method

A. Overview

We provide an overview of our proposed method via the training manner and pipeline, using a summer-to-winter (transforming summer domain to winter domain) image translation task as an example to describe our framework's details.

1) Training manner: In Fig. 3 (a), we use a summer image s and a winter image w from the Transient Attributes dataset [58] to extract the content codes (summer: C_s, winter: C_w) and style codes (summer: S_s, winter: S_w). E_s and E_w are the encoders of s and w, respectively. D_s and D_w are the decoders. By combining C_s with S_s to feed into D_s, we can reconstruct a summer image $s2s$. Similarly, by combining C_w with S_w to feed into D_w, a winter image $w2w$ can be reconstructed. The style codes Z_w and Z_s are randomly sampled from the normal distribution. By hypothesizing that Z_w is from the winter-style space and combining C_s with Z_w to feed into D_w, a winter image $s2w$ can be synthesized, as indicated by the red arrows. Similarly, we hypothesize that Z_s is from the summer-style space to combine with C_w to synthesize a summer image $w2s$ by D_s. For training, the above four processes are divided into cross-domain training ($s2w$ and $w2s$) and within-domain training ($s2s$ and $w2w$), which are deployed together [3].

2) Pipeline: In Fig. 3 (b), we utilize a pretrained panoptic segmentation network to obtain the panoptic perception of the input image scene. It provides the panoptic-level bounding boxes, categories, and masks to the corresponding modules

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

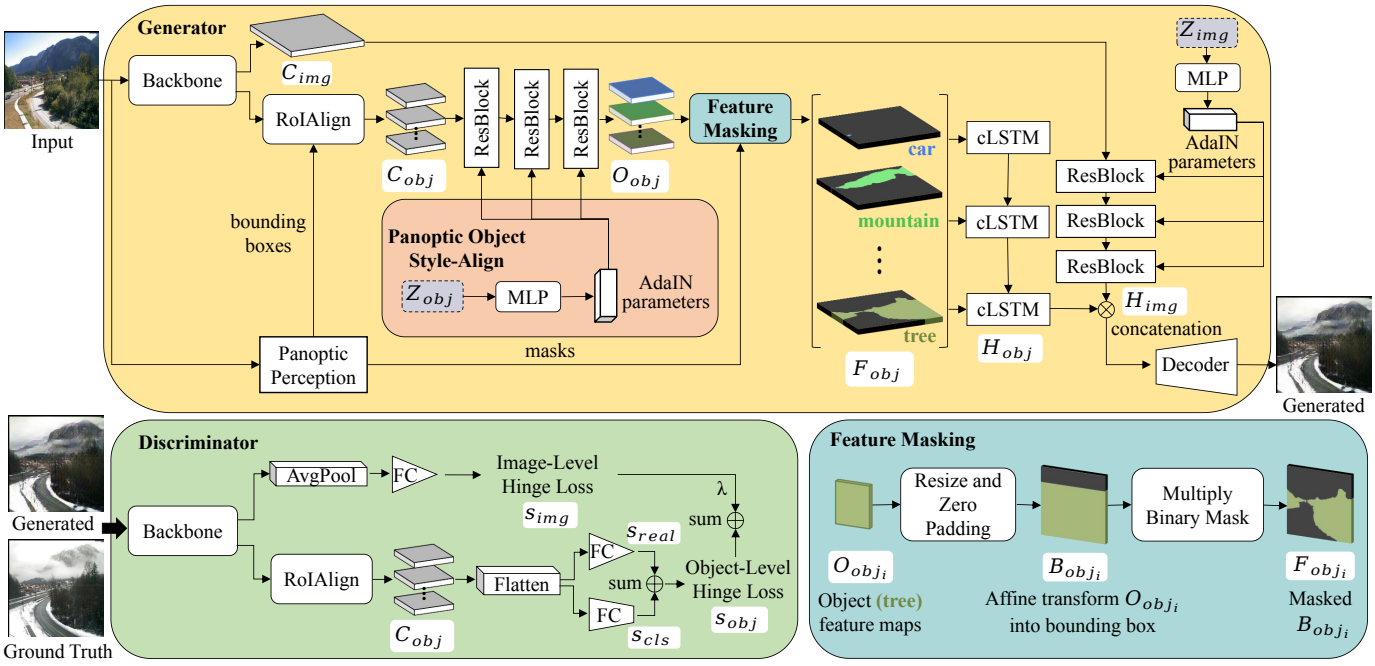


Fig. 4. Architecture. In the generator, the summer image is extracted by the backbone consisting of down-sampling residual blocks to image-level representations, which are cropped by RoIAlign by bounding boxes from the panoptic perception of the pretrained model to the panoptic-level representations. Both representations are aligned with sampled winter style codes from the normal distribution. The redundant background information of (panoptic-level) the style-aligned codes is removed by feature masking, and the results are reintegrated by cLSTM to combine with image-level representations for color image generation. In the discriminator, the translated image is extracted to the image-level and panoptic-level representations, which are processed to a fusion hinge loss consisting of an image-level realness score and object-level realness scores with category projection scores.

of the generator. They are provided to the basic RoIAlign [17] module and our proposed novel modules (panoptic object style-align and feature masking), respectively. The details are illustrated in the Architecture section. First, the image-level representation is extracted, the panoptic-level style codes Z_{obj} and image-level style codes Z_{img} are sampled from the normal distribution, and m is the number of objects perceived in the panoptic perception. Note that we treat both and 'thing' and 'stuff' as objects in the panoptic perception. Z_{obj} and Z_{img} are aligned with the corresponding object-level and image-level representations in the generator for a panoptic-level image translation. The translated images are fed into the discriminator, which calculates the object-level loss and image-level loss separately. Here, we utilize fusion hinge loss, which consists of the image and object adversarial hinge loss terms [59].

B. Architecture

Our architecture, as illustrated in Fig. 4, is built on a generator, a discriminator and the proposed novel modules (panoptic object style-align and feature masking). We deploy a generative adversarial learning setting via the summer and winter domain images from the Transient Attributes dataset [58] to illustrate our architecture.

1) Generator: In the generator, the input summer image s (e.g., 256×256) is extracted by a backbone module consisting of down-sampling residual blocks to obtain the image content codes C_{img} (size 32×32, dimension 256). Let $P = \{(category_i, bbox_i, mask_i)_{i=1}^{m}\}$ be the panoptic perception consisting of categories, bounding boxes, and masks, where m is the number of objects perceived from a pretrained panoptic segmentation network, and $category_i \in CAT$ (CAT defines 134 categories in the COCO-Panoptic dataset [18], here the 'thing' has 80 categories and the 'stuff' has 54 categories). C_{img} is cropped by RoIAlign [17] through the object bounding boxes of $P(bbox_i)_{i=1}^{m}$ into the object content codes $C_{obj} = \{C_{obj_i}\}_{i=1}^{m}$ (size 8×8, dimension 128). Define Z_{img} as the image-level style codes (dimension 256) and $Z_{obj} = \{Z_{obj_i}\}_{i=1}^{m}$ as the panoptic-level style codes (dimension 64), which are randomly sampled from the normal distribution; m is equal to the number of objects perceived in the panoptic perception. The goal of the generator in the summer-to-winter translation is to learn a generation function $G(\cdot)$, which is capable of translating summer image, s, to a generated winter image, $w^{'}$, via a given (Z_{img}, Z_{obj}):

$$w^{'} = G(s|Z_{img}, Z_{obj}; \Theta_G) \tag{1}$$

where Θ_G are the parameters of the generation function.

For the panoptic object style-align module, we use an MLP network to process the panoptic-level style codes $Z_{obj} = \{Z_{obj_i}\}_{i=1}^{m}$ to dynamically generate the parameters $y = (y_\gamma, y_\beta)$ of the adaptive instance normalization (AdaIN) [34] layers. Then, the object content codes C_{obj} are processed by the residual blocks with the AdaIN layers. The parameters of the AdaIN layers fuse the panoptic-level style with the content to translate the different objects in the target image.

$$AdaIN(x_i, y) = y_{\gamma,i}\left(\frac{x_i - \mu(x_i)}{\sigma(x_i)}\right) + y_{\beta,i} \tag{2}$$

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

where x_i is each feature map of C_{obj}, which is normalized separately, and then scaled and biased using the corresponding scalar components from style y. μ and σ are the channelwise means and standard deviation, respectively, and γ and β are the AdaIN parameters generated from Z_{obj}. This process achieves a panoptic object style-align, and we obtain the style-aligned object representations $O_{obj} = \{O_{obj_i}\}_{i=1}^{m}$.

$$O_{obj} = AdaIN(C_{obj}, Z_{obj}) \tag{3}$$

Similarly, the image-level style codes Z_{img} are also processed by the MLP network to generate the AdaIN parameters, which fuse the image-level style with the image content codes C_{img} by the residual blocks with the AdaIN layers to obtain a hidden representation H_{img}.

For the feature masking module, as illustrated in Fig. 4, O_{obj} contains m object feature maps $\{O_{obj_i}\}_{i=1}^{m}$. Since the object bounding boxes $P(bbox_i)_{i=1}^{m}$ define the size and location of each object in the original image, we first affine transform each object feature map O_{obj_i} into its corresponding original bounding box. Second, we perform zero-padding outside each bounding box in the image to obtain new object feature maps $B_{obj} = \{B_{obj_i}\}_{i=1}^{m}$. To remove the redundant background information outside the object contour, we further refine B_{obj} via the object masks $M = P(mask_i)_{i=1}^{m}$ for more precise object boundaries. Compared with the convolutional feature masking (CFM) layer [60] using the pixel projection method, after the affine transformation the size of each feature map in B_{obj} is the same as mask M. Therefore, we only need to align B_{obj} and M along the category sequence of $1 \sim m$ and multiply to mask the values outside of the object contour. Finally, we can obtain finer object feature maps $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

$$F_{obj} = B_{obj} \cdot M \tag{4}$$

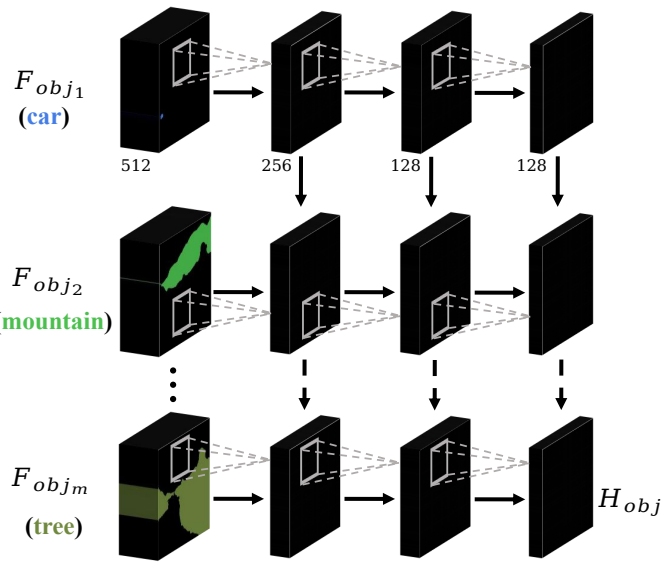


Fig. 5. Illustration of the convolutional long short-term memory (cLSTM) component. We use three layers of cLSTM to fuse all the object feature maps into hidden feature maps H_{obj}. The numbers of channels in each layer of cLSTM are 256, 128 and 128. The residual blocks are omitted for clarity. The first of each row is the object feature maps of $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

After feature masking, the masked object feature maps need to be fused into a well-hidden representation to generate a realistic target image. Therefore, we need to integrate all the objects into the desired locations and coordinate the object feature maps based on the other objects in the image. As shown in Fig. 5, the convolutional long short-term memory (cLSTM) [61] is a multilayer convolutional LSTM network, wherein the hidden states and cell states are both feature maps rather than vectors, which differs from the traditional LSTM [62]. The computation of the different gates is also performed by the convolutional layers. Therefore, cLSTM can better preserve the spatial information compared with the traditional vector-based LSTM. It can integrate each object feature map, $\{F_{obj_i}\}_{i=1}^{m}$, one-by-one, along the object sequence of $1 \sim m$ obtained by panoptic perception. The last output of cLSTM is used as the fused hidden representation H_{obj}. Different objects are sequentially fused while keeping their spatial locations in the image. We concatenate H_{obj} with H_{img} as H, which is upsampled by a decoder consisting of upsampled residual blocks to generate a translated winter image $w^{'}$.

2) Discriminator: As illustrated in Fig. 4, our discriminator consists of image-level and object-level classifiers. Similar to the generator, the translated image is encoded by the backbone as image content codes C_{img}, which are refined by RoIAlign [17] as object content codes $C_{obj} = \{C_{obj_i}\}_{i=1}^{m}$ by bounding boxes $P(bbox_i)_{i=1}^{m}$. The image-level classifier consists of global average pooling and a one-output fully connected (FC) layer to process C_{img} to obtain a scalar realness score s_{img}. The object-level classifier consists of a flattened layer and two FC layers. One FC layer processes C_{obj} to compute a realness score for each object, and is denoted by $s_{real} = \{s_{real_i}\}_{i=1}^{m}$. Another FC layer computes a category projection score [39], [63], [64] for each object, and is denoted by $s_{cls} = \{s_{cls_i}\}_{i=1}^{m}$, which is the inner product between a category embedding (transforming each category of $P(category_i)_{i=1}^{m}$ to a corresponding latent vector sampled from the normal distribution) and a linear projection (using an FC layer) of the downsampled C_{obj}. Therefore, the overall object-level loss of an object is $s_{obj_i} = s_{real_i} + s_{cls_i}$. The discriminator is denoted by $D(\cdot, \Theta_D)$ with the parameters Θ_D.

$$(s_{img}, s_{obj_i}, \cdots, s_{obj_m}) = D(I; \Theta_D) \tag{5}$$

Given an image I (ground truth w or generated $w^{'}$), the discriminator computes the prediction score for the image and the average scores for the objects.

C. Loss Function

The full objective of our model comprises three loss functions. We train the generator and discriminator networks end-to-end in an adversarial manner. The generator is trained to minimize the weighted sum of losses.

1) Adversarial Loss: We utilize the image-level and object-level fusion hinge version [59] of the standard adversarial loss [65] to train (Θ_G, Θ_D) in our PanopticGAN, which is utilized to ensure that each object's information can be optimized and to avoid losing too much information in the translation.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

$$l_k(I) = \begin{cases} \min(0, -1 + s_k); & \text{if } I \text{ is ground truth } w \\ \min(0, -1 - s_k); & \text{if } I \text{ is generated } w^{'} \end{cases} \quad (6)$$

where $k \in \{img, obj_i, \cdots, obj_m\}$. The overall loss is $l(I) = \lambda \cdot l_{img}(I) + \frac{1}{m}\sum_{i=1}^{m} l_{obj_i}(I)$ with the trade-off parameter λ (1.0 used in the experiment) in the fusion hinge losses between the image-level and the object-level. We define the losses for the discriminator and the generator [39],

$$\begin{aligned} L_{\text{adv}}(\Theta_D, \Theta_G) &= -\mathop{\mathbb{E}}_{(I)\ p_{all}(I)}[l(I)] \\ L_{\text{adv}}(\Theta_G, \Theta_D) &= -\mathop{\mathbb{E}}_{(I)\ p_{fake}(I)}[D(I; \Theta_D)] \end{aligned} \quad (7)$$

where minimizing $L(\Theta_D, \Theta_G)$ tries to tell the discriminator to distinguish the ground truth and the translated images, but minimizing $L(\Theta_G, \Theta_D)$ tries to fool the discriminator by translating the fine-grained images. $p_{all}(I)$ represents all of the ground truth and translated images, and $p_{fake}(I)$ represents the translated images.

2) Image Reconstruction Loss: We need $L_1^{img} = \left\| w^{'} - w \right\|_1$ to penalize the L_1 difference between the translated image $w^{'}$ and the ground truth w, and $\|_1$ calculates the L1 norm. Here, we mainly calculate the within-domain ($s2s$ and $w2w$) ways described in the training manner. Image reconstruction loss is essential for within-domain training and is also a very efficient pixelwise image generation optimization.

3) Perceptual Loss: We use L_{p} to alleviate the problem that the translated images are prone to producing distorted textures [16]. It is beneficial to keep the textures in a high-level space through the ReLU activation of the VGG-19 network [7],

$$L_{\text{p}} = \sum_k \frac{1}{C_k H_k W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} \left\| \phi_k(w^{'})_{i,j} - \phi_k(w)_{i,j} \right\|_1 \quad (8)$$

where $\phi_k(\cdot)$ represents the feature representations of the kth max-pooling layer in the VGG-19 network, and $C_k H_k W_k$ represents the size of the feature representations.

4) Full Objective: The final loss function is defined as:

$$L_{\text{total}} = \lambda_1 L_{\text{adv}} + \lambda_2 L_1^{img} + \lambda_3 L_{\text{p}} \quad (9)$$

where λ_i are the parameters balancing different losses. λ_i are usually the hyperparameters in the deep learning network trained through extensive experiments. We empirically determine the optimal tradeoff parameters after a large number of countless experiments.

D. Implementation Details

We provide the necessary explanation for some parameters set in the experiment, including the network architecture parameters and training parameters.

1) Network Architecture: The descriptions of the network architectures of PanopticGAN are shown in the Architecture section and we provide some detailed parameter explanations unmentioned in the main text. The backbone is composed of four residual blocks for downsampling. The RoIAlign operation crops from different sizes (32 and 16 in the experiment) image representations and combines them. We use a three-layer MLP network with spectral normalization and the leaky-ReLU activation function to process the sampled style codes. After a fusion of panoptic content and style, we obtain O_{obj} (category m, dimension 128, size 32×32). After resizing and zero padding, B_{obj} (category m, dimension 128, size 256×256) is multiplied by masks M (category m, size 256×256) to obtain F_{obj} (category m, dimension 128, size 256×256), which is processed by four downsampled convolution layers with a batch normalization. The result (category m, dimension 512, size 32×32) is fed into the cLSTM and then processed by six residual blocks to obtain H_{obj} (dimension 128, size 32×32), which is concatenated with H_{img} (dimension 128, size 32×32). H (dimension 256, size 32×32) is fed into the decoder, which consists of six residual blocks for upsampling and a final convolution layer with batch normalization and tanh activation.

2) Training: To stabilize the training of GANs, spectral normalization [66] is used for the layers in both the generator and the discriminator. For the activation function, we use leaky-ReLU with a slope of 0.2 in the modules. In L_{total}, the trade-off parameters, $\lambda_1 \sim \lambda_3$, are empirically set to 0.1, 1 and 10, respectively. Our model is trained using the Adam optimizer [67] with $\beta_1 = 0$ and $\beta_2 = 0.9$. The batch size is set to one for all the experiments. We set 400,000 iterations for training on four NVIDIA V100 GPUs. We train all models of the generator with a fixed learning rate of 10^{-4}, and train all models of the discriminator with a fixed learning rate of 0.005 until 200,000 iterations, and linearly decay up to 400,000 iterations. We also use a weight decay at a rate of 0.0001. The weights are initialized using the orthogonal initialization method [68].

IV. Experiments

We conduct extensive experiments to evaluate our method with competing I2I translation baseline tasks to show our superiority. We mainly evaluate three aspects, i.e., the image quality of the translated images, the object recognition performance of the translated images, and the visual odometry performance of the translated videos. The evaluation of the image quality mainly conducts a comprehensive comparison based on three aspects, realism, sharpness and diversity. For the object recognition performance, we mainly perform panoptic segmentations on the translated images and compared the panoptic quality (PQ) [18] series' metric values. Because panoptic segmentation includes semantic segmentation for the background, and instance segmentation for the foreground, it can be evaluated more comprehensively than instance segmentations and object detections. For the visual odometry (VO) performance, we validate the performance of the direct VO method, namely, the direct sparse odometry (DSO) [20].

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

Specifically, we choose six subsequences from the entire route. Each subsequence includes several characteristics, such as multiple corners and a straight-shaped route. Note that there were no overlapping segments between the training sequences and the testing sequences. Based on the specific metrics, we compare our method against several baselines, both qualitatively and quantitatively. The comparisons include the image quality of the translated images, the object recognition performance on translated images and the visual odometry performance on translated videos.

A. Baselines

For the baselines, we mainly select the competing methods, which achieved good results, as the state-of-the-art baselines to compare with our method. In addition, we aimed to show our proposed panoptic-level method as an upgraded image-to-image translation method that is better than the instance-level and image-level methods. Therefore, we summarize the baselines into image-level and instance-level baselines to facilitate the evaluation comparisons. For the competing methods, MUNIT [3], BicycleGAN [1], TSIT [30], FDIT [31] and pSp [33] were categorized under image-level I2I translations. As SCGAN [8] uses saliency maps for salient regions perception, we regarded SCGAN as an instance-level translation. INIT [2] and InstaFormer [45] are instance-level methods, so we also implemented them as instance-level evaluations for fairer comparisons. Since the experimental results of TICCGAN [25] showed high competitiveness in the visual odometry performance evaluation, we added it as one of the baselines for evaluating comparisons of the visual odometry performances. To achieve an adequately fair comparison, we added the panoptic perception to image-level competing methods, i.e., MUNIT, BicycleGAN, TSIT, FDIT and pSp, as well as TICCGAN. It was extracted from a pretrained panoptic segmentation network [19]. The panoptic perception (as an additional feature channel) was concatenated with the image feature maps for training. Additionally, to neutralize the advantage brought by the additional pretrained panoptic network, to provide a adequately fair comparison, we also added panoptic-replaced instance-level competing methods (SCGAN+Panoptic, INIT+Panoptic, InstaFormer+Panoptic), that is, instances were replaced with panoptic information (instance + semantics) based on the pretrained panoptic network for the training and result evaluation comparisons.

B. Datasets

We trained and evaluated our model on the Transient Attributes [58] and KAIST-MS [21] datasets for day-to-night, summer-to-winter and thermal-to-color I2I translation tasks using 256×256 resolution images. In the day-to-night I2I translation task, we used 17,823 images for training and 2,287 images for evaluation; in the summer-to-winter I2I translation task, the training set had 17,674 images and the evaluation set had 2558 images; in the thermal-to-color I2I translation, the training set consisted of 11,610 images, and the evaluation set had 2,541 images. Our method used panoptic-level perception to avoid losing too much information in the translation. For panoptic perception in the training and inference processes of the day-to-night and summer-to-winter I2I translation tasks, we utilized a Panoptic FPN model [19] pretrained on the COCO-Panoptic dataset [18] to perceive from the input day images and summer images. For panoptic perception in the training process of the thermal-to-color I2I translation task, we perceived it from the paired color images via the pretrained Panoptic FPN model on the COCO-Panoptic dataset; in the inference, it is perceived from the input images via the pretrained Panoptic FPN model on a compact dataset of thermal panoptic segmentation, and the source data were the pairs of thermal and color images from the partial KAIST-MS [21] dataset. The unaugmented source data from our contributed dataset contained 2,026 pairs of thermal and color images based on the partial KAIST-MS [21] dataset, which was annotated via the Segments.ai platform for panoptic segmentation annotation by three professionally trained annotators. The annotated datasets could be augmented by various image manipulations for a variety of tasks. We show an overview of the annotated dataset via this link[1]. Please refer to the insights section on the overview tab for the distribution of the categories and the number of annotated objects ('thing' and 'stuff'). For the annotation quality and the details of images, we show the samples of the dataset on paired color images via this link[2].

C. Evaluation Metrics

We evaluated the image quality and object recognition performance of the translated images, and evaluated the visual odometry performance of the translated videos. We chose the Human Preference (HP), Inception Score (IS) [69], Fréchet Inception Distance (FID) [70] and Diversity Score (DS) metrics instead of the Peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) [71] metrics to evaluate image quality. For images generated by the GAN learning models, the traditional PSNR and SSIM metrics deviate from human visual perception [72]. Panoptic Quality (PQ) [18] series metrics were used to evaluate the object recognition performance. The absolute pose error [73] metric was used for the visual odometry performance.

1) Human Preference (HP): HP is a user perceptual study that compares the image quality of translated images from different methods through human cognition. Twenty participants (average age: 30.20, std: 20.46) were shown the translated results from the evaluation set corresponding to all three different tasks (thermal-to-color, day-to-night and summer-to-winter I2I translation). They were generated by different methods with the original input images and the ground truth images grouped together for comparison. We divided each evaluation set results into five groups to show to five persons among the 20 participants in turn. Each group's images are selected with the best three results corresponding to realism, object sharpness and scene similarity with unbiased weights (1:1:1) to ensure an adequate fairness of the evaluation. The total number of selections was calculated as a final comprehensive percentage score.

[1]**Overview:** https://segments.ai/panoptic/visible/
[2]**Samples:** https://segments.ai/panoptic/visible/samples

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

TABLE I
HUMAN PREFERENCE, INCEPTION SCORE, FRÉCHET INCEPTION DISTANCE AND DIVERSITY SCORE METRICS ARE USED FOR THE IMAGE QUALITY EVALUATIONS OF THE THERMAL-TO-COLOR (T2C), DAY-TO-NIGHT (D2N) AND SUMMER-TO-WINTER (S2W) I2I TRANSLATION TASKS. HIGHER HUMAN PREFERENCE, HIGHER INCEPTION SCORE, HIGHER DIVERSITY SCORE AND LOWER FRÉCHET INCEPTION DISTANCE INDICATE BETTER IMAGE QUALITY.

Method	Human Preference ↑			Inception Score ↑			Fréchet Inception Distance ↓			Diversity Score ↑		
	t2c	d2n	s2w	t2c	d2n	s2w	t2c	d2n	s2w	t2c	d2n	s2w
MUNIT	2.40%	1.23%	3.36%	2.29	1.50	1.92	98.51	98.79	93.97	0.46	0.65	0.62
BicycleGAN	1.10%	3.14%	2.41%	2.61	1.86	1.81	98.82	97.96	92.23	0.47	0.60	0.61
TSIT	3.22%	10.01%	3.35%	2.64	1.78	1.96	95.38	80.86	81.32	0.43	0.67	0.64
SCGAN	1.23%	6.37%	1.59%	2.59	1.62	1.58	96.83	92.40	86.45	0.39	0.53	0.49
SCGAN+Panoptic	–	–	–	2.62	1.68	1.67	91.21	90.67	83.20	0.42	0.57	0.53
INIT	9.51%	2.24%	11.87%	2.70	1.22	1.84	83.26	76.73	78.90	0.37	0.65	0.57
INIT+Panoptic	–	–	–	2.75	1.53	1.86	79.34	70.82	74.67	0.45	0.66	0.61
FDIT	16.07%	5.93%	8.74%	2.63	1.59	1.90	90.31	72.65	74.23	0.49	0.68	0.60
pSp	11.34%	20.49%	14.78%	2.73	1.68	1.91	79.08	73.86	76.27	0.52	0.70	0.63
InstaFormer	24.68%	18.12%	23.92%	2.80	1.89	1.95	74.71	72.34	73.52	0.51	0.69	0.64
InstaFormer+Panoptic	–	–	–	2.82	1.90	1.98	73.64	70.83	72.19	0.50	0.70	0.67
Ours	**30.45%**	**32.47%**	**29.98%**	**2.85**	**1.93**	**2.01**	**72.73**	**69.40**	**71.16**	**0.54**	**0.72**	**0.69**

2) Inception Score (IS): IS is a popular metric that measures the quality of the generated images from GANs. It uses an Inception V3 network pretrained on the ImageNet-1000 classification benchmark and computes a statistics score of the network's outputs [74],

$$IS(G) \approx exp\{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{1000} p(y=j|I_i) log\frac{p(y=j|I_i)}{\hat{p}(y=j)}\} \quad (10)$$

where N synthesized images I_i are translated from the generator model G, $\hat{p}(y=j) = \frac{1}{N}\sum_{i=1}^{N} p(y=j|I_i)$. The two desirable qualities of the image synthesis are evaluated by IS, meaning the synthesized images should contain clear and meaningful objects, and the diverse images from the different categories in ImageNet should be observed in the synthesized images.

3) Fréchet Inception Distance (FID): FID improved the inception score (IS) [69] by incorporating statistics from real images. Similar to IS, FID also uses an Inception V3 network pretrained on ImageNet to compute the fréchet distance [75] between two Gaussian distributions fitted to the synthesized images and real images [74]. Therefore, the lower the FID is, the better a generator model is. The fréchet distance is defined by,

$$FID^2((\mu^0, \sum^0), (\mu^1, \sum^1)) = \left\|\mu^0 - \mu^1\right\|_2^2 + T_r(\sum^0 + \sum^1 - 2(\sum^0\sum^1)^{1/2}) \quad (11)$$

where $(\mu^0, \sum^0)$ and $(\mu^1, \sum^1)$ denote the mean vector and the covariance matrix of the Gaussian distribution fitted on the synthesized and real images, respectively.

4) Diversity Score (DS): DS measures the differences between the paired images generated from the same input by computing the perceptual similarity in deep feature space [37]. We used the LPIPS metric [72] for diversity scoring, and the pretrained AlexNet [76] for feature extraction.

$$DS(I_1, I_2) = \sum_{i=1}^{n}\frac{1}{|\Lambda_i|}\sum_{p\in\Lambda_i}\left\|w_i \odot (x_1^i(p) - x_2^i(p))\right\|_2^2 \quad (12)$$

where n layers of unit-normalized features (in channel dimension) x^i are extracted, $|\Lambda_i|$ denotes the spatial area of a feature map, and w_i denotes the learned parameters.

5) Panoptic Quality (PQ): PQ was adopted to evaluate the object recognition performance, and PQ combines segmentation quality (SQ) and recognition quality (RQ) [18],

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP}\text{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \quad (13)$$

where SQ sums all of the intersection over union (IoU) ratios for the true positives (TP) and evaluates how closely matched the predicted segments are with their ground truths. RQ is a blend of precision and recall, where all TPs, half False-Positives (FP), and False-Negatives (FN) are divided. It combines precision and recall to identify how effective a trained model is at getting a prediction right. PQ^{Th}, SQ^{Th}, and RQ^{Th} are used on 'thing' (Th) categories; PQ^{St}, SQ^{St}, and RQ^{St} are used on 'stuff' (St) categories. PQ series metrics combine mean IoU (mIoU) in SQ and average precision (AP) in RQ for a more comprehensive score than the instance segmentations and object detections. Furthermore, since our framework was based on panoptic perception, using PQ series metrics was more appropriate than the traditional object recognition metrics.

6) Absolute Pose Error (APE): APE was adopted to evaluate the visual odometry performance. It consists of the median absolute translational error t_{rel} and absolute rotational error r_{rel} as proposed in the KITTI Odometry benchmark [73]. We ran the DSO [20] method and calculated the APE scores to evaluate the trajectory similarity between the original frame sequences (video) and the corresponding translated frame sequences (video) from different competing approaches.

D. Qualitative Results

For image quality, the human preference results in Table I show that our approach achieved significantly higher scores in the human perceptual study of the different summer-to-winter, day-to-night and thermal-to-color I2I translation tasks compared with the other approaches. Fig. 6 demonstrates

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

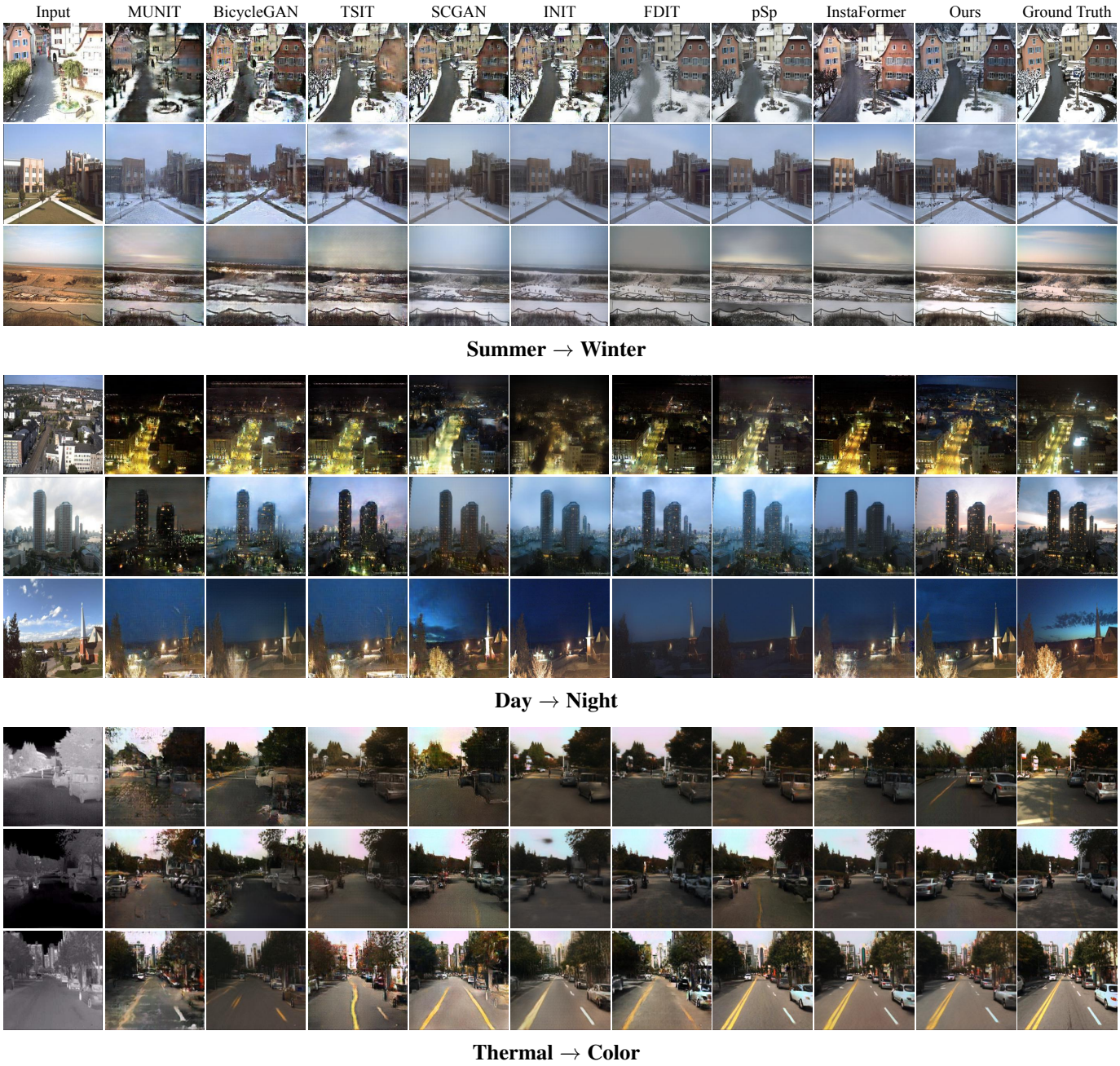


Fig. 6. Comparisons of the image quality of the translated images from different methods with the input and ground truth images. The results of our method tend to generate more realistic and fine-grained images, and show tiny objects in more detail.

that our PanopticGAN could translate to higher fidelity and brightly colored images and show tiny objects in more detail. In contrast, the results from other methods were more blurry, distorted and missing small objects. For the translated objects, our results tend to have better sharpness, more natural color and display a high diversity (e.g., the appearance of cars). On the other methods side, the object sharpness was not satisfactory, the style was far from the ground truth and there was insufficient diversity. Moreover, Fig. 7 also shows a contrast in the details of the generated objects between our results and the competing methods. This demonstrated the superiority of our method on translated objects, producing sharp boundaries and adequate coloring, and maintaining a certain diversity, e.g., the types and styles of cars. In the image quality comparison of Fig. 6, although the results of the InstaFormer method are more prominent in color contrast, its object boundaries are not consistent with the ground truth, and the detailed textures of the objects are not as good as our results.

For the object recognition performance, we used the panoptic segmentation result from the pretrained Panoptic FPN model on the COCO-Panoptic dataset. We only showed the object recognition results on the thermal-to-color I2I translation task, because the translated night images from the day-to-night I2I translation and the winter images from the summer-to-winter I2I translation tasks had the disadvantage

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

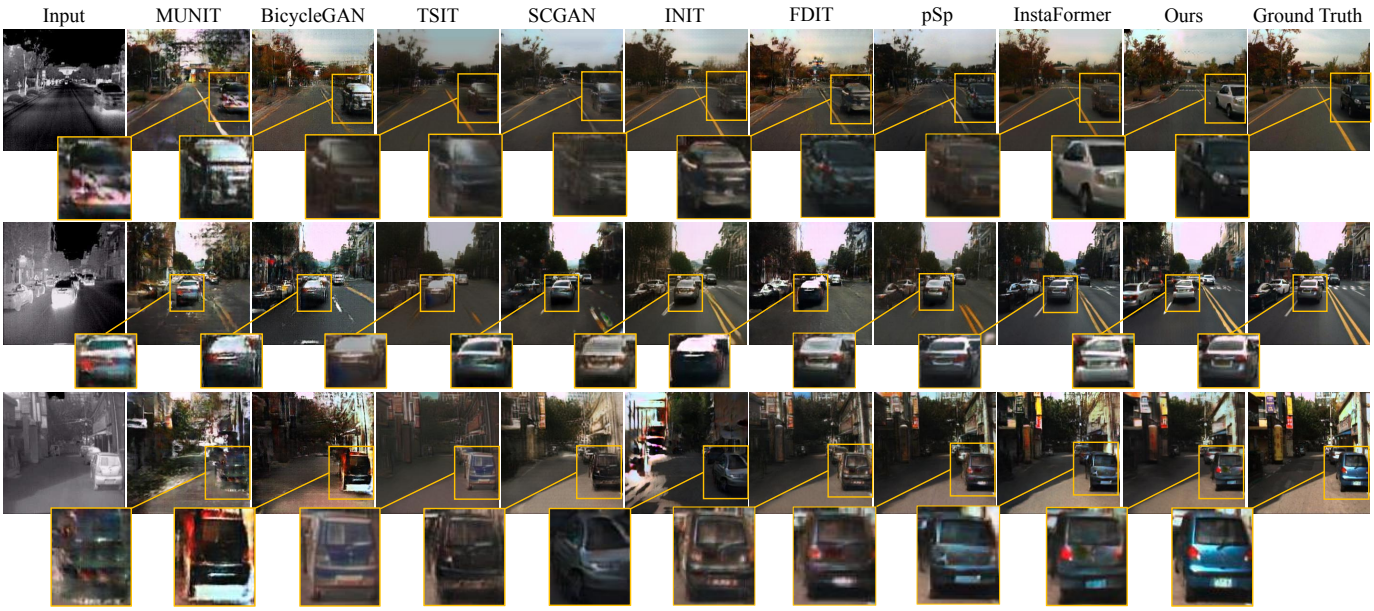


Fig. 7. Comparison results on the details of the translated objects from the different approaches with the ground truth images. The results of our method tend to have better sharpness, a more natural color, and display high diversity.

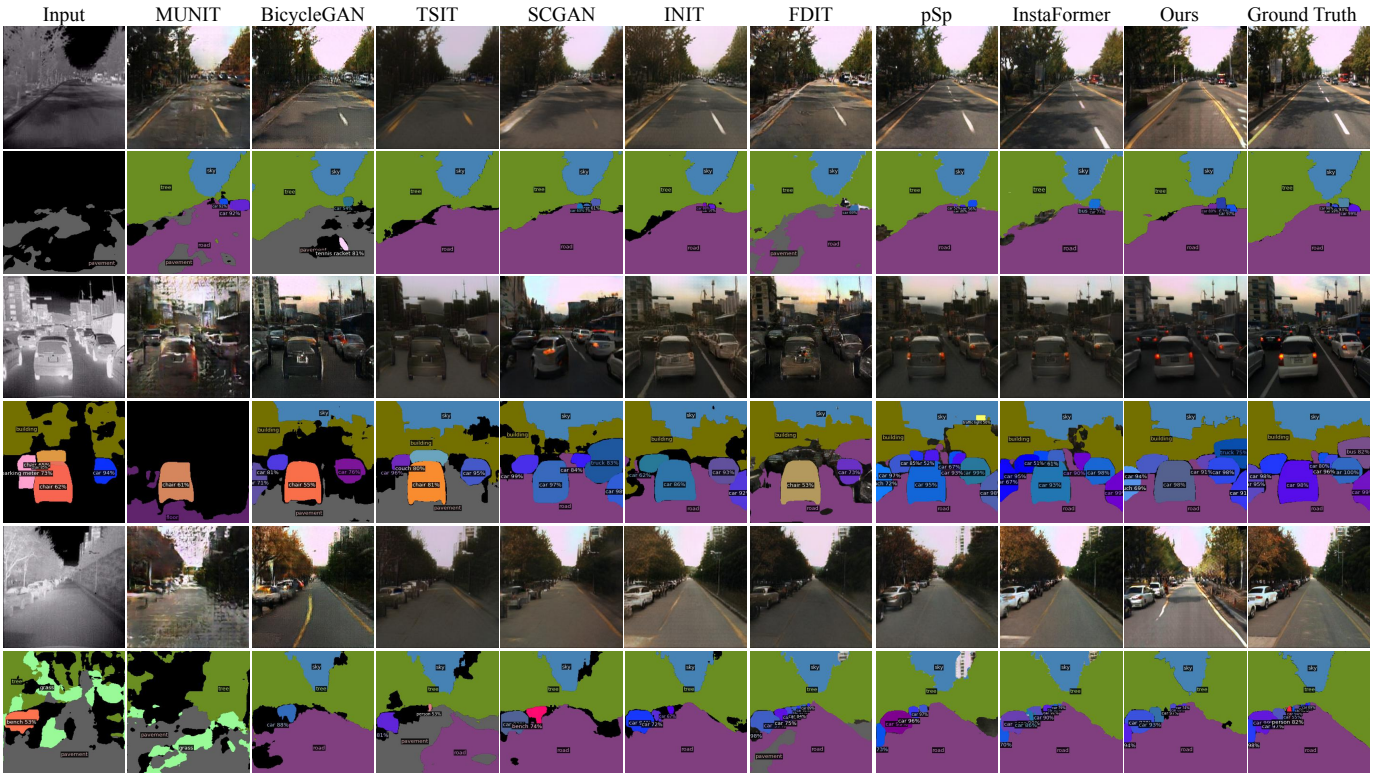


Fig. 8. The object recognition performance of the translated images from different approaches. In each group, the upper row shows the translated images, and the lower row shows the results from the corresponding panoptic segmentation. Our method tends to achieve better results, especially for tiny objects.

of insignificant differences for an object recognition comparison. As illustrated in Fig. 8, we show the panoptic segmentation comparison results with the baselines. The results demonstrated that our method could achieve a better object recognition performance, e.g., the number and boundaries of the cars, structure of the sky, tree and road, and the areas with relatively fewer recognition failures. Compared with the recognition results of the original thermal image, the results of our translated color images were significantly improved. This verified the advantages of our method when adapted for image enhancement.

For the visual odometry (VO) performance, we added TICCGAN [25] to the baselines since its results showed a

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

TABLE II

THE PANOPTIC QUALITY (PQ) CORRESPONDS TO PANOPTIC RECOGNITION ACCURACY, THE SEGMENTATION QUALITY (SQ) CORRESPONDS TO MEAN INTERSECTION OVER UNION (MIOU) AND THE RECOGNITION QUALITY (RQ) CORRESPONDS TO AVERAGE PRECISION (AP). THESE ARE UTILIZED TO COMPREHENSIVELY EVALUATE THE PANOPTIC-LEVEL OBJECT RECOGNITION PERFORMANCE OF THE TRANSLATED IMAGES FROM THE DIFFERENT APPROACHES IN THE THERMAL-TO-COLOR I2I TRANSLATION TASK. THE PQ^{TH}, SQ^{TH}, AND RQ^{TH} METRICS ONLY CONSIDER 'THING' (TH) CATEGORIES; THE PQ^{ST}, SQ^{ST}, AND RQ^{ST} METRICS ONLY CONSIDER 'STUFF' (ST) CATEGORIES. (HIGHER IS BETTER)

Method	PQ ↑	SQ ↑	RQ ↑	PQ^{Th} ↑	SQ^{Th} ↑	RQ^{Th} ↑	PQ^{St} ↑	SQ^{St} ↑	RQ^{St} ↑
MUNIT [3]	3.3	12.1	4.2	0.6	9.6	0.8	9.0	17.5	11.3
BicycleGAN [1]	4.3	16.8	5.5	0.8	13.1	1.2	10.9	23.9	13.6
TSIT [30]	6.4	17.2	8.1	2.1	13.3	3.3	13.9	26.4	15.3
SCGAN [8]	5.6	15.2	7.4	1.7	11.8	2.6	13.6	22.5	17.4
SCGAN+Panoptic	6.3	16.7	7.8	2.5	12.4	3.7	14.2	24.3	18.8
INIT	7.2	19.6	9.0	3.1	15.4	3.9	16.7	29.1	20.9
INIT+Panoptic	7.7	20.4	9.8	3.3	15.9	4.4	17.3	29.8	21.2
FDIT	7.4	20.1	9.6	3.4	14.7	4.3	17.1	29.6	20.3
pSp	7.8	20.7	10.2	3.8	16.1	4.6	17.6	30.2	20.7
InstaFormer	7.9	21.3	10.8	3.9	16.8	4.8	17.9	30.7	21.2
InstaFormer+Panoptic	8.1	22.0	11.1	**4.2**	17.1	5.0	18.2	30.9	21.3
Ours	**8.3**	**22.7**	**11.3**	**4.2**	**17.5**	**5.1**	**18.4**	**31.0**	**21.6**

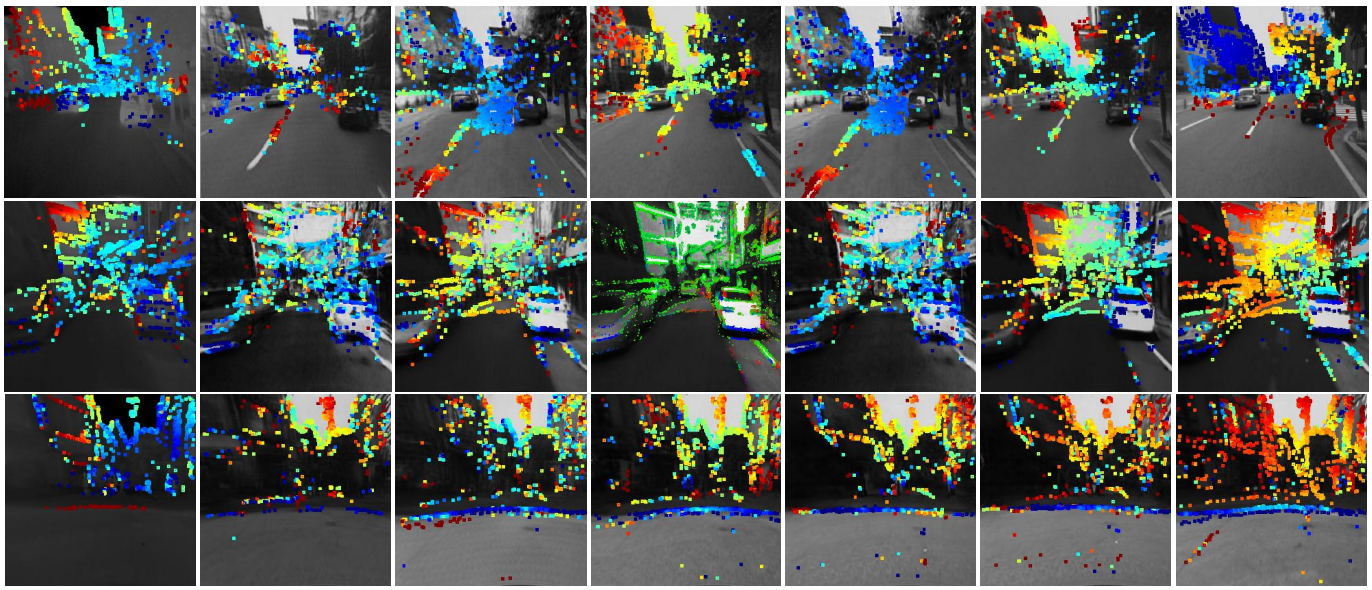


Fig. 9. The key point comparisons of the translated frames in the visual odometry for the different approaches. The blue points indicate a large depth value; the red points indicate a small value; and the green points indicate the initialized value. Our results have more robust points and are more consistent with the ground truth.

superior performance. Therefore, we selected four methods (i.e., TICCGAN [25], INIT [2], pSp [33] and InstaFormer [45]) with the highest performance to compare with our results. We ran the DSO [20] method on both the original thermal frame sequences (video) and the corresponding translated color frame sequences through different competing approaches. First, Fig. 9 shows the comparison results of the extracted key points with the depth maps from the different methods. The translated frames from our method had more key points than the baselines, and they could be robustly obtained and are more consistent with the ground truth. They facilitated more accurate depth estimations and made the tracking robust against the brightness variation in the environment for better 3D reconstructions (3D point clouds) and estimated scene movements (trajectory). Second, Fig. 10 shows that our method could obtain a better VO performance in the trajectory, e.g., compared with the results from the other methods, the trajectories predicted by our method were closer to the ground-truth trajectories. Note that we measured the absolute trajectory error because the raw KAIST-MS dataset [21] was published with some discontinuous video clips, and the test sequences were relatively short, i.e., less than 50 m. As seen in the different subgraphs, our results still consistently showed that the predicted trajectories matched the expected orientations and poses in the subsequences compared with the other methods.

E. Quantitative Results

For the image quality, the inception score, fréchet inception distance and diversity score in Table I show that our approach achieved superiority in the image quality of the translated images compared with other approaches. Overall, our method

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

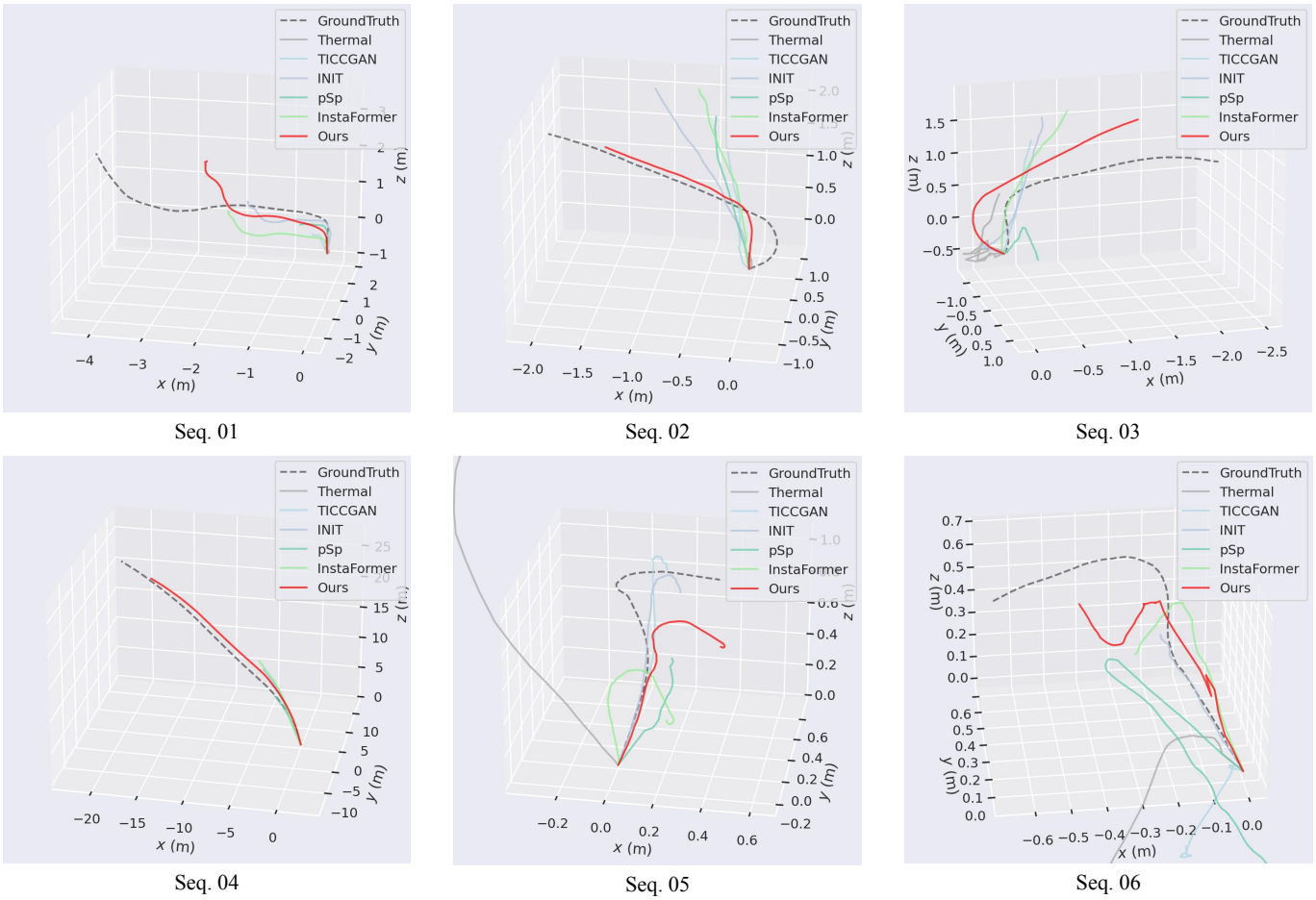


Fig. 10. The performance comparisons of the translated subsequences (video) in the visual odometry for the different approaches, mainly for the trajectories. Our estimated trajectories tend to be more consistent with the ground truth.

TABLE III
THE T_{REL} AND R_{REL} METRICS (LOWER IS BETTER) ARE UTILIZED TO EVALUATE THE VISUAL ODOMETRY PERFORMANCE (TRAJECTORY).

Seq.	Thermal		TICCGAN		INIT		INIT+Panoptic		pSp		InstaFormer		InstaFormer+Panoptic		Ours	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
01	2.30	2.00	1.24	1.57	1.05	1.56	1.01	1.32	1.41	1.60	1.07	1.24	0.96	1.20	**0.89**	**1.16**
02	2.45	1.81	1.75	2.79	1.59	2.86	1.37	2.23	1.45	1.97	1.38	1.71	1.31	1.57	**1.28**	**1.49**
03	2.37	3.14	1.31	1.29	1.20	1.38	1.14	1.27	1.17	1.26	1.12	1.20	1.09	1.17	**1.03**	**1.10**
04	2.21	2.42	2.03	1.89	1.34	1.60	1.07	1.13	1.08	1.14	0.97	1.01	0.56	0.78	**0.32**	**0.45**
05	3.30	2.56	2.44	2.23	2.18	2.46	2.09	1.98	2.23	2.17	2.14	2.09	2.10	1.83	**2.05**	**1.71**
06	6.51	7.02	7.41	7.52	4.05	3.17	2.99	3.02	6.21	5.53	3.24	3.36	2.65	2.87	**2.16**	**2.34**

outperformed the baselines since we were able to avoid losing too much information in the translation. The higher inception score and lower fréchet inception distance of our method verified that the images generated by our method had higher fidelity and sharper object information. The higher diversity score demonstrated that our method was more flexible and highly robust when a scene was invariant, especially for the objects generated on the image.

For the object recognition performance, Table II shows that the object recognition results from our method earned state-of-the-art scores compared with the competing trained models on all PQ, SQ, RQ, PQ^{Th}, SQ^{Th}, RQ^{Th}, PQ^{St}, SQ^{St}, and RQ^{St} object recognition metrics. From the score differences, the table data demonstrates that our results were uniformly higher than the state-of-the-art competing methods by a certain distance, which confirms the superiority of our method.

For the visual odometry performance, since DSO is greatly affected by the gradient points through the boundaries of objects, our translation results with high sharpness benefited the performance of running DSO [20] on the translated videos. As expected, Table III validated that our method could obtain a better performance on monocular visual odometry. The six rows of the table correspond to the comparison of the predictions produced by running DSO with the ground truth trajectories for the different videos shown in Fig. 10. We measured the absolute trajectory error (APE) between the predicted videos and the ground truth videos. The APE value of each sequence showed that our absolute translational error

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

TABLE IV
THE ABLATION STUDY OF OUR MODEL BY REMOVING THE DIFFERENT LOSSES AND MODULES. L_{obj}: OBJECT-LEVEL HINGE LOSS; L_1^{img}: IMAGE RECONSTRUCTION LOSS; L_{p}: PERCEPTUAL LOSS; M_{masking}: FEATURE MASKING MODULE; M_{panoptic}: PANOPTIC OBJECT STYLE-ALIGN MODULE; M_{clstm}: CLSTM MODULE. THE INCEPTION SCORE (IS), FRÉCHET INCEPTION DISTANCE (FID) AND DIVERSITY SCORE (DS) ARE TO EVALUATE THE IMAGE QUALITY; THE PANOPTIC QUALITY (PQ), SEGMENTATION QUALITY (SQ) AND RECOGNITION QUALITY (RQ) ARE TO EVALUATE THE OBJECT RECOGNITION PERFORMANCE; THE ABSOLUTE TRANSLATIONAL ERROR ($\mathrm{T_{REL}}$) AND ABSOLUTE ROTATIONAL ERROR ($\mathrm{R_{REL}}$) METRICS ARE TO EVALUATE THE VISUAL ODOMETRY PERFORMANCE.

Metrics	w/o L_{obj}	w/o L_1^{img}	w/o L_{p}	w/o M_{masking}	w/o M_{panoptic}	w/o M_{clstm}	Full Model
IS	2.24	2.64	2.66	2.53	2.44	2.63	**2.85**
FID	110.4	104.3	97.1	101.6	120.7	103.2	**72.7**
DS	0.47	0.45	0.42	0.47	0.43	0.42	**0.54**
PQ	5.3	6.4	6.7	6.1	5.9	6.8	**8.3**
SQ	16.2	20.4	19.4	19.8	18.4	16.7	**22.7**
RQ	7.7	10.1	10.5	10.2	9.1	11.1	**11.3**
$\mathrm{t_{rel}}$	2.03	1.43	1.50	1.41	2.07	1.81	**1.28**
$\mathrm{r_{rel}}$	2.10	1.63	1.60	1.67	1.93	1.78	**1.37**

($\mathrm{t_{rel}}$) and the rotational error ($\mathrm{r_{rel}}$) were lower than the others, which was consistent with the estimated trajectory comparison results in Fig. 10. Especially for the frame sequences in discontinuous environments, our subset error consistently remained the lowest, which demonstrated the robustness of our method in an environmental scene adaptation. In contrast, the results from the other methods were not very stable, especially the result from the TICCGAN method in set 06, which was even lower than the original thermal result.

F. Ablation Study

We demonstrate the necessity of all the losses and modules of our proposed model by comparing the inception score (IS) [69], fréchet inception distance (FID) and diversity score (DS) [37] for the image quality; panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ) for the object recognition performance [18]; and absolute translational error ($\mathrm{t_{rel}}$) and absolute rotational error ($\mathrm{r_{rel}}$) for the visual odometry performance. They were deployed for the experimental results using several ablated versions of our model trained on the KAIST-MS [21] dataset of the thermal-to-color translation task. As shown in Table IV, removing any losses decreased the overall performance. Removing L_{obj} and L_1^{img} has a lower IS and DS and a higher FID due to generating low fidelity images and objects with fewer variations; the PQ, SQ, and RQ were all significantly decreased because L_{obj} computed the category projection scores for the objects; the $\mathrm{t_{rel}}$) and $\mathrm{r_{rel}}$ values rose because the DSO method was affected by the gradient points through the boundaries of the objects. The low sharpness of the objects also decreased the visual odometry performance. By removing L_{p}, the model could not alleviate the problem that translated images are prone to producing distorted textures. This inevitably decreased the image quality, object recognition performance and visual odometry performance. Removing the M_{masking}, M_{panoptic} or M_{clstm} module decreased the overall performance, which demonstrates their necessity. This was because M_{masking} sharpens object boundaries and M_{clstm} sequentially integrates different objects back into the image. In particular, removing M_{panoptic} destroyed the entire foundation of our proposed panoptic-level framework, and the overall performance was significantly decreased. Therefore, the above study results of the losses and modules showed the reasonability of our model design.

G. Model Efficiency

Model efficiency can influence practical applications, and is mainly measured by the computational cost, model complexity and processing time. The model complexity includes the time complexity (number of model operations) and space complexity (number of model parameters). The time complexity determines the training and prediction time of the model, and can be measured in floating-point operations (FLOPs). It is defined separately as a computational cost in our paper. The space complexity determines the number of parameters of the model. Due to the curse of dimensionality limitation, the more parameters in the model, the larger the amount of data required to train the model. It can be measured in parameters (Params), and we separately defined it as the model complexity in our paper. For the processing time, we calculated the average processing time (Avg PT) per image for the different networks of the competing baselines. As shown in Table V, we present the model efficiency comparisons between our scores and the best baselines' scores. Here, we list the competing baselines, TSIT [30], INIT [2], INIT+Panoptic, pSp [33], InstaFormer [45] and InstaFormer+Panoptic. Table V shows that the Params of our model were lower than those of the other models, and the FLOPs were only slightly higher than those of the TSIT and InstaFormer models. For the different I2I translation tasks (summer-to-winter, day-to-night and thermal-to-color), our model spent less processing time on average than the other models. Overall, our method outperformed the baselines, since we used panoptic-level perception to avoid losing too much information in the translation. Moreover, our model did not incur substantial computational costs or model complexity, and the average processing time kept it competitive.

H. Failure Cases

As shown in Fig. 11, the cars ('thing') generated by our model have better sharpness, more natural color, and display diversity. However, the generated road ('stuff') has an incorrect texture, i.e., largely different lane markings and zebra crossings compared with the InstaFormer [45] method.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

TABLE V
THE FLOATING-POINT OPERATIONS (FLOPS) AND PARAMETERS (PARAMS) EVALUATE THE COMPUTATIONAL COST AND MODEL COMPLEXITY. THE AVERAGE PROCESSING TIME (AVG PT) PER IMAGE EVALUATES THE PROCESSING SPEED; THE LOWER, THE BETTER.

Model	Params (M)	FLOPs (G)	Avg PT (ms) t2c	d2n	s2w
TSIT	116.1	**50.8**	19.1	19.7	**18.4**
INIT	130.3	62.9	22.3	21.0	21.6
INIT+Panoptic	141.6	68.3	22.7	22.3	21.9
pSp	124.7	56.3	18.7	19.2	20.1
InstaFormer	116.8	51.0	17.9	20.4	20.5
InstaFormer+Panoptic	120.3	53.4	18.2	21.4	21.1
Ours	**113.6**	51.4	**17.6**	**19.0**	19.9

Fig. 11. An example of the failure cases from the proposed PanopticGAN. In extreme cases, the high diversity of panoptic objects may result in incorrect textures being generated to large semantic regions.

Although our method refined each panoptic ('thing' + 'stuff') object during the translations to achieve a higher-fidelity generation, it also allows each object to have high diversity. In extreme cases, if a 'stuff' object with a large area is generated with incorrect textures and there is a significant difference with the ground truth, it may result in a decrease in the overall image quality. Additionally, it may impact the consistency of the continuous frames to further affect the visual odometry performance. The 'stuff' normally has a larger region than the 'thing' and therefore needs to be considered with the entire image context to avoid producing incorrect textures on large areas in extreme cases. These challenges will be considered in our future work.

V. CONCLUSION

In this paper, we presented a pioneering panoptic-level I2I translation method (PanopticGAN). Compared to the image-level and instance-level methods, PanopticGAN refines the object semantics on both the foreground and background to avoid the semantic information loss. To sharpen the object boundaries, we proposed a feature masking module to extract the object-specific representations along the contours. PanopticGAN can generate higher-fidelity images/videos to further enhance the performances of the object recognition and visual odometry tasks. Furthermore, we annotated a panoptic segmentation dataset for the thermal-to-color translation. The experimental results showed the superiority of our proposed method. In future work, we aim to overcome the limitation that the high diversity of panoptic objects may result in incorrect textures being generated to large semantic regions in extreme cases. To this end, we plan to explore introducing an attention mechanism to consider the entire image context in the translation. In addition, we also plan to investigate and extend our proposed method to address the image translations between SAR (synthetic aperture radar) and optical modalities, the subsequent SAR image interpretation, as well as the SAR ship detection, e.g., utilizing enhanced subsequences from SAR-to-optical image translation to improve ship tracking performance represents a valuable and meaningful challenge.

REFERENCES

[1] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Multimodal image-to-image translation by enforcing bi-cycle consistency," in *Advances in neural information processing systems*, 2017, pp. 465–476.

[2] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3683–3692.

[3] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[4] D. S. Tan, Y.-X. Lin, and K.-L. Hua, "Incremental learning of multi-domain image-to-image translations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1526–1539, 2020.

[5] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.

[6] R. Chen and Y. Zhang, "Learning dynamic generative attention for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8368–8382, 2022.

[7] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[8] Y. Zhao, L.-M. Po, K.-W. Cheung, W.-Y. Yu, and Y. A. U. Rehman, "Scgan: Saliency map-guided colorization with generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[9] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1418–1430, 2021.

[10] L. Fu, H. Yu, F. Juefei-Xu, J. Li, Q. Guo, and S. Wang, "Let there be light: Improved traffic surveillance via detail preserving night-to-day transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[11] Z. Zhu, Y. Meng, D. Kong, X. Zhang, Y. Guo, and Y. Zhao, "To see in the dark: N2dgan for background modeling in nighttime scene," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 492–502, 2020.

[12] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 155–170.

[13] E. Jung, N. Yang, and D. Cremers, "Multi-frame gan: image enhancement for stereo visual odometry in low light," in *Conference on Robot Learning*. PMLR, 2020, pp. 651–660.

[14] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[18] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[19] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[20] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[21] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[22] L. Zhang, P. Ratsamee, B. Wang, Z. Luo, Y. Uranishi, M. Higashida, and H. Takemura, "Panoptic-aware image-to-image translation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 259–268.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[24] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.

[25] X. Kuang, J. Zhu, X. Sui, Y. Liu, C. Liu, Q. Chen, and G. Gu, "Thermal infrared colorization via conditional generative adversarial network," *Infrared Physics & Technology*, vol. 107, p. 103338, 2020.

[26] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[28] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[29] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *arXiv preprint arXiv:1907.10830*, 2019.

[30] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "Tsit: A simple and versatile framework for image-to-image translation," in *European Conference on Computer Vision*. Springer, 2020, pp. 206–222.

[31] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang, "Frequency domain image translation: More photo-realistic, better identity-preserving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 930–13 940.

[32] Y. Pang, J. Xie, and X. Li, "Visual haze removal by a unified generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3211–3221, 2018.

[33] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296.

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[35] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4561–4569.

[36] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1219–1228.

[37] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8584–8593.

[38] T. Sylvain, P. Zhang, Y. Bengio, R. D. Hjelm, and S. Sharma, "Object-centric image generation from layouts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2647–2655.

[39] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 531–10 540.

[40] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," *arXiv preprint arXiv:1812.10889*, 2018.

[41] S. Ma, J. Fu, C. W. Chen, and T. Mei, "Da-gan: Instance-level image translation by deep attention generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5657–5666.

[42] L. Jiang, M. Xu, X. Wang, and L. Sigal, "Saliency-guided image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 509–16 518.

[43] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7968–7977.

[44] T. Chen, W. Xiong, H. Zheng, and J. Luo, "Image sentiment transfer," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4407–4415.

[45] S. Kim, J. Baek, J. Park, G. Kim, and S. Kim, "Instaformer: Instance-aware image-to-image translation with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 321–18 331.

[46] Y. Lin, Y. Wang, Y. Li, Y. Gao, Z. Wang, and L. Khan, "Attention-based spatial guidance for image-to-image translation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 816–825.

[47] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," *IEEE Transactions on Multimedia*, vol. 23, pp. 1654–1665, 2021.

[48] A. Dundar, K. Sapra, G. Liu, A. Tao, and B. Catanzaro, "Panoptic-based image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8070–8079.

[49] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[51] X. Sun, C. Chen, X. Wang, J. Dong, H. Zhou, and S. Chen, "Gaussian dynamic convolution for efficient single-image segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2937–2948, 2021.

[52] X. Weng, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Stage-aware feature alignment network for real-time semantic segmentation of street scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[53] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[54] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang, "Ssap: Single-shot instance segmentation with affinity pyramid," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 642–651.

[55] X. Zhang, H. Li, F. Meng, Z. Song, and L. Xu, "Segmenting beyond the bounding box for instance segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 704–714, 2021.

[56] X. Wang, C. Shen, H. Li, and S. Xu, "Human detection aided by deeply learned semantic masks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2663–2673, 2019.

[57] Q. Chen, A. Cheng, X. He, P. Wang, and J. Cheng, "Spatialflow: Bridging all tasks for panoptic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2288–2300, 2020.

[58] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on graphics (TOG)*, vol. 33, no. 4, pp. 1–11, 2014.

[59] J. H. Lim and J. C. Ye, "Geometric gan," *arXiv preprint arXiv:1705.02894*, 2017.

[60] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3992–4000.

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

[62] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3288547

[63] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[64] T. Miyato and M. Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[66] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[69] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[70] S. Ravuri and O. Vinyals, "Classification accuracy score for conditional generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[72] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

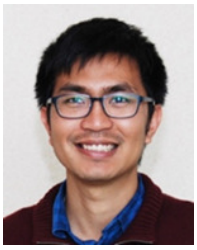
[73] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[74] W. Sun and T. Wu, "Learning layout and style reconfigurable gans for controllable image synthesis," *arXiv preprint arXiv:2003.11571*, 2020.

[75] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

Liyun Zhang (Graduate Student Member, IEEE) received the M.S. degree in computer technology engineering from the School of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Technology, Osaka University, Osaka, Japan. His research interests include computer vision, robotics, multimodal learning, and embodied reasoning.



Photchara Ratsamee (Member, IEEE) received his M.E. and Ph.D. degrees from the Graduate school of Engineering Science, Osaka University in 2012 and 2015, respectively. He was an assistant professor at Cybermedia Center, Osaka University from 2015-2021. Currently, he is an associate professor (lecturer) at the Faculty of Robotic and Design, Osaka Institute of Technology (OIT). His research interests include Robot Vision, Human Robot Interaction, Augmented Reality for Robot and Hybrid Robot.

Zhaojie Luo (Member, IEEE) received his M. Eng. and Dr. Eng. degrees from Kobe University in Japan. From 2019 to 2020, he has been a researcher at the Department of Electrical & Computer Engineering of National University of Singapore (NUS). He is currently an assistant professor at the Department of Information and Communications Technology, Osaka University. His research interests include multimedia signal processing, facial expression recognition, speech synthesis, image signal processing and statistical signal processing. He is a member of IEEE, ISCA and ASJ. He has published more than 30 publications in major journals and international conferences, such as IEEE-T-ASLP, IEEE Trans. Multimedia, EURASIP JASMP, INTERSPEECH, SSW, ICME, ICPR, *etc.*

Yuki Uranishi (Member, IEEE) is currently an an associate professor of Cybermedia Center, Osaka University, Japan. He received his M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2005 and 2008, respectively. His research interests include Computer Vision, Augmented Reality and Human-Computer Interaction. Contact him at yuki.uranishi.cmc@osaka-u.ac.jp.

Manabu Higashida received his B.S. in Mathematics from Tokyo Institute of Technology in March 1989, M.S. in Mathematics from Tokyo Institute of Technology in March 1991, and Ph.D. in Engineering from Osaka University in 1994. He has been a lecturer at the Information Media Education and Research Division, Cybermedia Center, Osaka University since October 2014.

Haruo Takemura (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Osaka University, Osaka, Japan, in 1982, 1984, and 1987, respectively. He was a Researcher at Advanced Telecommunication Research Institute, International (ATR-I), Kyoto, Japan, from 1987 to 1994, and an Associate Professor at Nara Institute of Science and Technology, Nara, Japan. He has been a Professor at the Infomedia Education Division, Cybermedia Center, Osaka University, since 2001. He is in charge of campus wide deployment of learning management system (LMS) and other IT systems for education. His research interests include interactive computer graphics, human–computer interaction, mixed reality, and their applications in education, including learning analytics.

Authorized licensed use limited to: OSAKA UNIVERSITY. Downloaded on October 01,2023 at 14:06:54 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.