Integrating Panoptic-Level to Image Translation for Enhanced Robotic Perception

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2024

Zhang LIYUN

Thesis Committee

Prof. Haruo Takemura (Osaka University)
Prof. Yoshinobu Kawahara (Osaka University)
Prof. Yasushi Yagi (Osaka University)
Assoc. Prof. Yuki Uranishi (Osaka University)

List of Publications

Peer-Reviewed Journals

1. Zhang. L., Ratsamee, P., Luo, Z., Uranishi, Y., Higashida, M., Takemura, H. (2023). Panoptic-Level Image-to-Image Translation for Object Recognition and Visual Odometry Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). (Early Access)*

2. Zhang. L., Liu, N., Hou, Y., Liu, X. (2014). Uneven Illumination Image Segmentation Based on Multi-threshold S-F. *Opto-Electronic Engineering Vol.41 No. pp. 81-87. (OEE).*

Peer-Reviewed Conference Proceedings

1. Zhang. L., Ratsamee, P., Uranishi, Y., Higashida, M., Takemura, H. (2022). Thermal-to-Color Image Translation for Enhancing Visual Odometry of Thermal Vision. *IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR).*

2. Zhang. L., Ratsamee, P., Wang, B., Luo, Z., Uranishi, Y., Higashida, M., Takemura, H. (2023). Panoptic-aware Image-to-Image Translation. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).*

Abstract

Addressing the challenges of low resolution and blurry characteristics in thermal images which are used in robotic vision perception leads to an unsatisfactory performance, this research focuses on enhancing visual odometry and object recognition tasks through innovative image-to-image translation. This allows rescue robots with thermal cameras penetrating thick smoke or dust in disaster environments to acquire high-quality color scene images translated from low-quality thermal vision, to explore 3D mapping, object detection, and rescuing in incipient disaster environments.

To overcome this challenge, this thesis proposed a novel generative adversarial network (PanopticGAN) for panoptic-aware image-to-image translation, which is capable of converting low-resolution and blurry thermal images into high-resolution and finer-grained color images. The generated images are further used in visual SLAM and object recognition tasks to greatly improve performance to meet practical needs.

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 and videos are further used.

Our proposed PanopticGAN network as a panoptic-level image-to-image translation 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 and videos are used in.

We also annotate a compact panoptic segmentation dataset for the thermal-to-color image translation task. Extensive experiments are conducted to demonstrate the effectiveness of our PanopticGAN over the latest methods.

Keywords: Incipient disaster environments, Rescue robot, 3D mapping, Thermal-to-color image translation, Panoptic-level Image-to-image translation, Image enhancement, Visual odometry, Object recognition.

Acknowledgements

In 2020, I entered the Integrated Media Environment Lab (Takemura lab) at Osaka University as a research student and continued to apply as a doctoral course student in a next following semester. I have received a lot of support in many aspects from the lab members. So, I would like to thank the following people for supporting me during my study.

First, I would like to thank my dissertation committee. Prof. Tatsuhiro Tsuchiya, Prof. Haruo Takemura, Prof. Masanori Hashimoto and Assoc. Prof. Yuki Uranishi. Your detailed feedback and advice have been very important to me. Also, my dissertation cannot be done without the committee's comments and review.

I am indebted to my research group leader, Assist. Prof. Photchara Ratsamee for being my guide and encouraging me through the complicated barrier of the AI (artificial intelligence) field with patience and kindness. I would also like to thank Prof. Haruo Takemura, my supervisor, Assoc. Prof. Manabu Higashida and Assoc. Prof. Jason Orlosky for the valuable comments and advice.

I am also thankful for all present, and past members of Takemura Lab, and friends who have taken their time to help me with the working environment, daily kinds of stuff, and inspiring activities.

Special thanks to Sysmex, Ltd., for the research grant and special researcher position that supported me in carrying on my research projects.

Finally, to my parents and my family for all the understanding and generous opportunity to give a chance for me to pursue my dreams and future of AI.

Zhang Liyun
Osaka University
March 2024

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Robot perception involves the ability of robotic systems to measure and understand their environment using various sensors, such as cameras, lidar, radar, and force sensors. Due to the variability of environments, robots often operate in diverse and unpredictable settings. For instance, autonomous driving in low-light or smoke-filled disaster environments presents challenges where the presence of unforeseen obstacles can significantly affect perception accuracy.

To address these challenges, there is a growing need for advanced image translation techniques using machine learning. Image translation refers to the process of transforming images from one representation or style to another, which can be crucial for enhancing robot perception. For example, in dark or smoke-filled environments, a robot equipped with a thermal camera can see through darkness or smoke. However, the low resolution of thermal images can make robot perception unreliable. Translating low-quality thermal images into high-quality color images can significantly improve a robot's ability to perceive its environment. This advancement is particularly useful for more advanced processing techniques like Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM), which are crucial in such conditions. They can be flexibly used in robotics, AR/VR, and autonomous driving applications.

1.1 Thermal Vision for the Perception of Rescue Robots in Disaster Scenes

Thermal infrared cameras can capture invisible heat radiation emitted or reflected by objectives to generate thermal images, regardless of lighting condi-

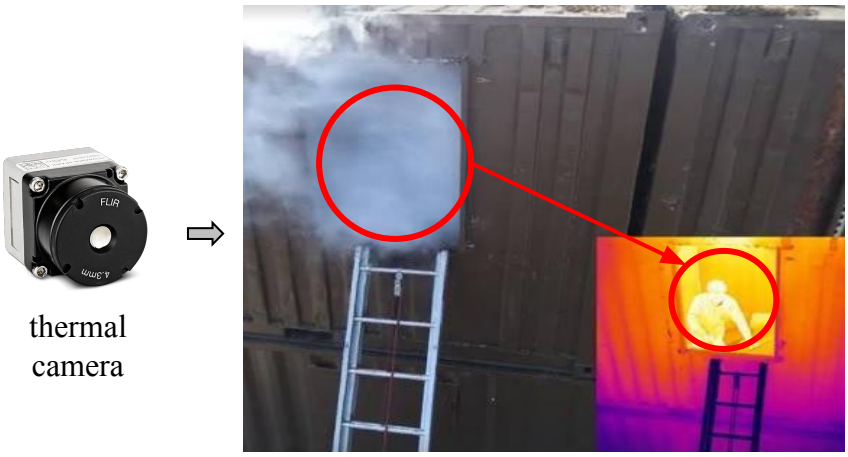
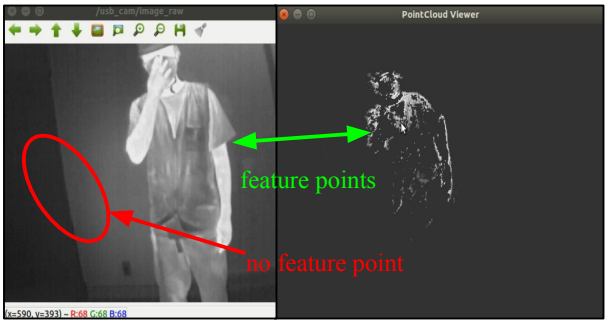


Figure 1.1. Thermal camera can capture the objective in a smoke-filled environment. The thermal camera is a BOSON 320 4.3mm Lens from FLIR Corporation, US. The smoke-filled image was taken by a DJI drone during a drill in Menlo Park, Calif.

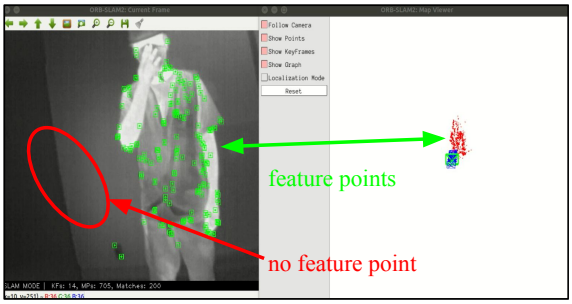
tions. This allows thermal cameras to penetrate smoke or fog and recognize objectives well in the dark. As illustrated in Figure 1.1, a normal camera in a fire scene full of smoke is barely visible from the situation in the window; the bottom right side is a thermal image, which can penetrate the smoke to capture the objective, this is greatly beneficial to the rescue work in disasters. Therefore, thermal images have been gradually valued in computer vision fields, e.g., Hwang et al.[3] applied thermal images in object recognition, and Khattak et al.[4] deployed simultaneous localization and mapping (SLAM) [5] on thermal image frames.

However, these works also further discovered the existing problems in thermal images, i.e., low resolution and blurry attributions [6], which result in unsatisfactory performance in the SLAM. However, because both feature-based methods [7, 8] and direct methods [9, 10] rely on image gradient-based key point extraction, which provides fewer high-quality points in thermal vision. Moreover, the low resolution and blurry attributions [6] of thermal images result in unsatisfactory performance in VO compared with color images, i.e., the estimated trajectory deviates significantly from Ground Truth.

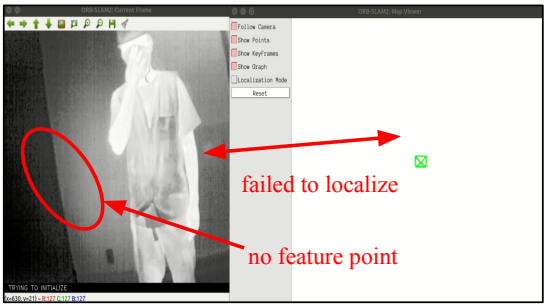
As illustrated in Figure 1.2, we conducted a comparative test analysis of the



(a) LSD-SLAM



(b) ORB-SLAM2



(c) ORB-SLAM2 after histogram equalization

Figure 1.2. Existing problems of SLAM on thermal frames. The thermal images were taken by a thermal camera at Kobe Fire Academy, Hyogo, Japan.

(a) Thermal images

(b) Object recognition results

Figure 1.3. Panoptic segmentation methods are used for object recognition results on thermal images.

performance of thermal frames directly in some popular SLAM applications. As illustrated in Figure 1.2(a), the result of LSD-SLAM [9] shows that recognized feature points are less and there is no feature point on background; in Figure 1.2(b), the result of ORB-SLAM2 [7] on thermal frames also shows that recognized feature points are less and only can be recognized on objective but no feature point on the background; in Figure 1.2(c), the result of ORB-SLAM2 after histogram equalization for enhancement show that there is no feature point recognized on objective and background.

At the same time, the performance of object recognition on thermal images is also unsatisfactory compared with color images. We also utilized the panoptic segmentation [11] model for the segmentation and object recognition in thermal images. As shown in Figure 1.3, the results of object recognition were not satisfactory; a number of objects were either not recognized or misidentified.

1.2 Thermal Image-to-Color Image Translation

Inspired by the multi-frame GANs [12] that transformed night images into day images to improve resolution and alleviate the blurry problem for an impressive gain on SLAM task. We aim to transform low-resolution and blurry thermal images into fine-grained color images to obtain a better performance in SLAM or visual odometry (VO) tasks compared with using thermal images in visual odometry directly. As shown in Figure 1.4, we proposed the thermal-to-color image translation model that converts blurry thermal frames to fine-grained color frames, which can obtain the improved semi-dense map in monocular visual odometry for positioning and search.

Also, transforming low-resolution and blurry thermal images into fine-grained color images (termed thermal-to-color) and using them in object recognition can greatly improve the performance of object recognition tasks compared with using thermal images in object recognition directly.

Image-to-image (I2I) translation is a challenging problem in the computer vision field. From the disentangled representation perspective [13, 14, 15], first, the input image representation needs to be extracted as content code and

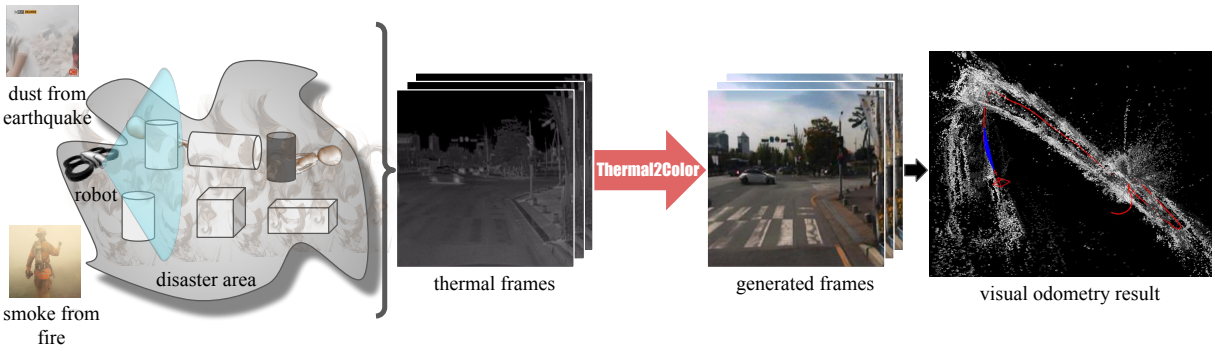


Figure 1.4. The application of thermal-to-color image translation for thermal vision enhancement in the visual odometry task.

combined with the style code sampled from the color style space for image translation. Translated images need to retain the content information of the input domain image and obtain the style of the target domain [14, 16]. Several tasks can be considered I2I translation problems, e.g., superresolution [17], [18], neural style transfer [19], and colorization [20, 21]. 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.[22] utilized a night-to-day image translation to enhance the object detection performance at nighttime; Zhu et al.[23] proposed a night-to-day image translation for the performance enhancement of background modeling; ForkGAN [24] presented a task-agnostic image translation to boost multiple vision tasks; MFGAN [12] transferred night videos to day videos to improve the resolution for impressive gains in the visual odometry and simultaneous localization and mapping (SLAM) [5] 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

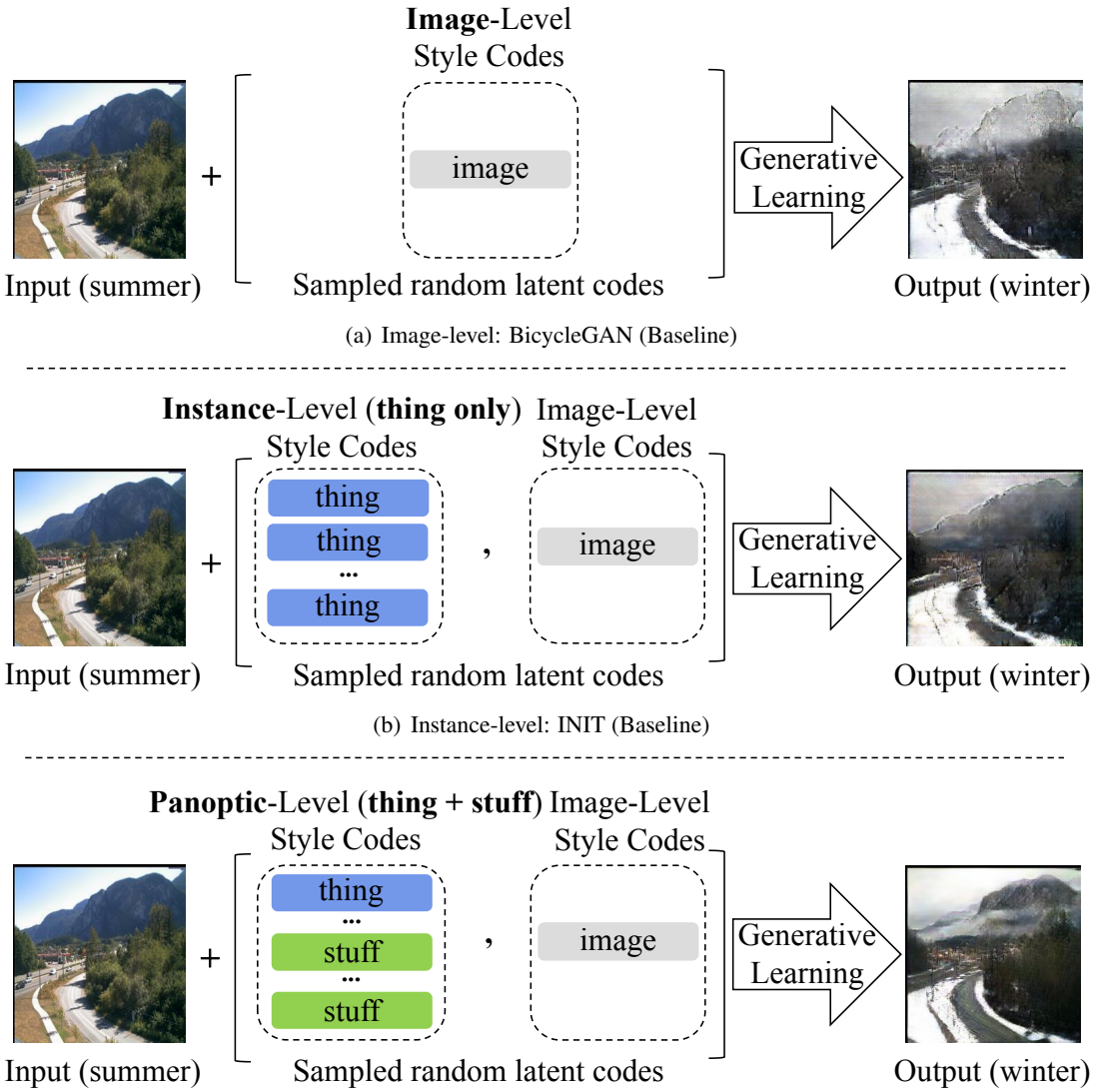
problem and an important challenge.

I2I translation has evolved from image-level methods [25, 26] to instance-level methods [2], resulting in significant improvements in the fidelity of translated images. As illustrated in Fig. 1.5 (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. 1.5 (b) uses a pretrained instance segmentation network [27] to extract the instance perception from the input image, which contains object bounding boxes, categories, and masks. It uses Region of Interest Align (RoIAlign) [27] 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.

1.3 Panoptic-level Image to Image Translation

In contrast, panoptic-level I2I translations can thoroughly refine the foreground object instances 'thing' and background semantic regions 'stuff' [11] 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.

The author proposed PanopticGAN uses a pretrained panoptic segmentation [28] network to extract the panoptic perceptions. In Fig. 1.5 (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.



(a) Image-level: BicycleGAN (Baseline)

(b) Instance-level: INIT (Baseline)

(c) Panoptic-level: PanopticGAN (Proposed)

Figure 1.5. 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.

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.

1.4 Statement of the Problem

- In order to realize a panoptic-level thermal-to-color image translation model and further apply it to enhance visual odometry and object recognition performance, we need to propose a novel generative model at the panoptic level. This involves acquiring panoptic-level image perception to generate high-quality images, using the generated images for object recognition tasks, and comparing the generated continuous frame videos with visual odometry tasks for experimental verification. Due to these reasons, the challenges of studying this method can be divided into four research questions respectively.

1.5 Research Questions

- How to design a universal model at the panoptic level that can transform between different modality images and generate high-quality target images.
- How to construct an efficient module that can ensure the accurate extraction and transformation of the information of each object during the generation process, and maintain high-quality object information in the generated results to support the benefits of enhanced object recognition and visual odometry performance.
- How to utilize the high-quality results from the panoptic-level image translation model for performance enhancement experiments in object recognition and visual odometry tasks respectively.

- How to validate the superiority of the proposed model and the evaluation criteria to assess the benefits of enhanced object recognition and visual odometry.

1.6 Philosophy

Considering the research question from the last section, the author has set up philosophies to answer the research question for the panoptic-level image translation model for performance enhancement of object recognition and visual odometry tasks. This philosophy consists of four main topics respectively.

- First, 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.
- Second, 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.
- Third, 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.
- Fourth, 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. We also annotated a compact panoptic segmentation dataset on a partial KAIST-MS dataset [3] for the thermal-to-color translation task.

1.7 Discussion

The presence of various dangerous elements such as fires or aftershocks often prevents rescue personnel from immediate intervention. In such situations, deploying rescue robots with enhanced thermal vision capabilities could play a pivotal role. These robots could swiftly navigate the disaster area, create 3D maps, and detect survivors, thereby bolstering rescue efforts and potentially saving more lives.

In disaster sites with dust or smoke-filled environments, robots equipped with enhanced vision can deal with extremely low visibility more effectively. Particularly, the monocular direct method VO, which utilizes image sequence frames acquired by RGB cameras, can achieve motion trajectory estimation and semi-dense 3D reconstruction. This approach reduces the reliance on extra sensors. However, tracking the RGB camera pose in poorly illuminated conditions (e.g., night scenes) or in the presence of airborne obscurants (e.g., dust, fog, and smoke) remains a challenge for current VO methods. In contrast, thermal cameras (e.g., Long Wave Infrared sensors) present a viable alternative, offering visibility through these challenging conditions.

The background of this dissertation is the use of rescue robots to replace human rescue teams in environmental mapping and rescue in the early stages of disasters, but the main goal is to improve the visual perception of robots equipped with thermal cameras. That is, the panoptic-level image translation from low-quality thermal images to high-quality color images is used to enhance visual odometry and object recognition performance.

When using thermal images captured by thermal cameras for 3D mapping, we will meet the few feature points and bad localization caused by the low resolution and blur characteristics of thermal images in SLAM application. Therefore, we use the panoptic-level image translation model to convert the original thermal image into a highly fine-grained color image, which will improve the problem of feature points and localization, thereby enhancing the performance of visual odometry.

However, although we utilize panoptic perception to achieve the generation

of higher-quality color images, it also makes each object highly diverse. It is possible to generate large-area objects with incorrect textures, resulting in a decrease in overall image quality and mapping performance. These challenges will also be considered in our future work.

1.8 Outline of the Dissertation

The overall concept of this dissertation is shown in Figure 1.6. This dissertation consists of five chapters, each of the chapters will be briefly described below.

- ✓ Chapter 1, Introduction, this chapter starts with the background on thermal vision for the perception of rescue robots in disaster scenes and image-to-image translation model-based thermal vision enhancement. This chapter explains the challenges, research questions, philosophy, and contribution of the proposed method.
- ✓ Chapter 2, Literature Review, this chapter reviews the previous works of image-to-image translation, instance-level image-to-image translation, panoptic-level image-to-image translation, and enhancement for visual odometry (VO).
- ✓ Chapter 3, Panoptic-level image-to-image translation, this chapter presents the concept of a pioneering and universal panoptic-level image translation network model, explains its advantages, and elaborates on the model's training manners and structural design.
- ✓ Chapter 4, Feature masking for enhancement of object recognition and visual odometry, this chapter introduces a novel feature masking module, explains its advantages integrated panoptic-level I2I translation network to sharpen object boundaries and translate higher fidelity objects, and to enhance the performance of object recognition and visual odometry. It also presents a detailed explanation of core modules and experimental results.
- ✓ Chapter 5, Conclusion, this chapter summarizes the major findings and outlines suggestions for future work.

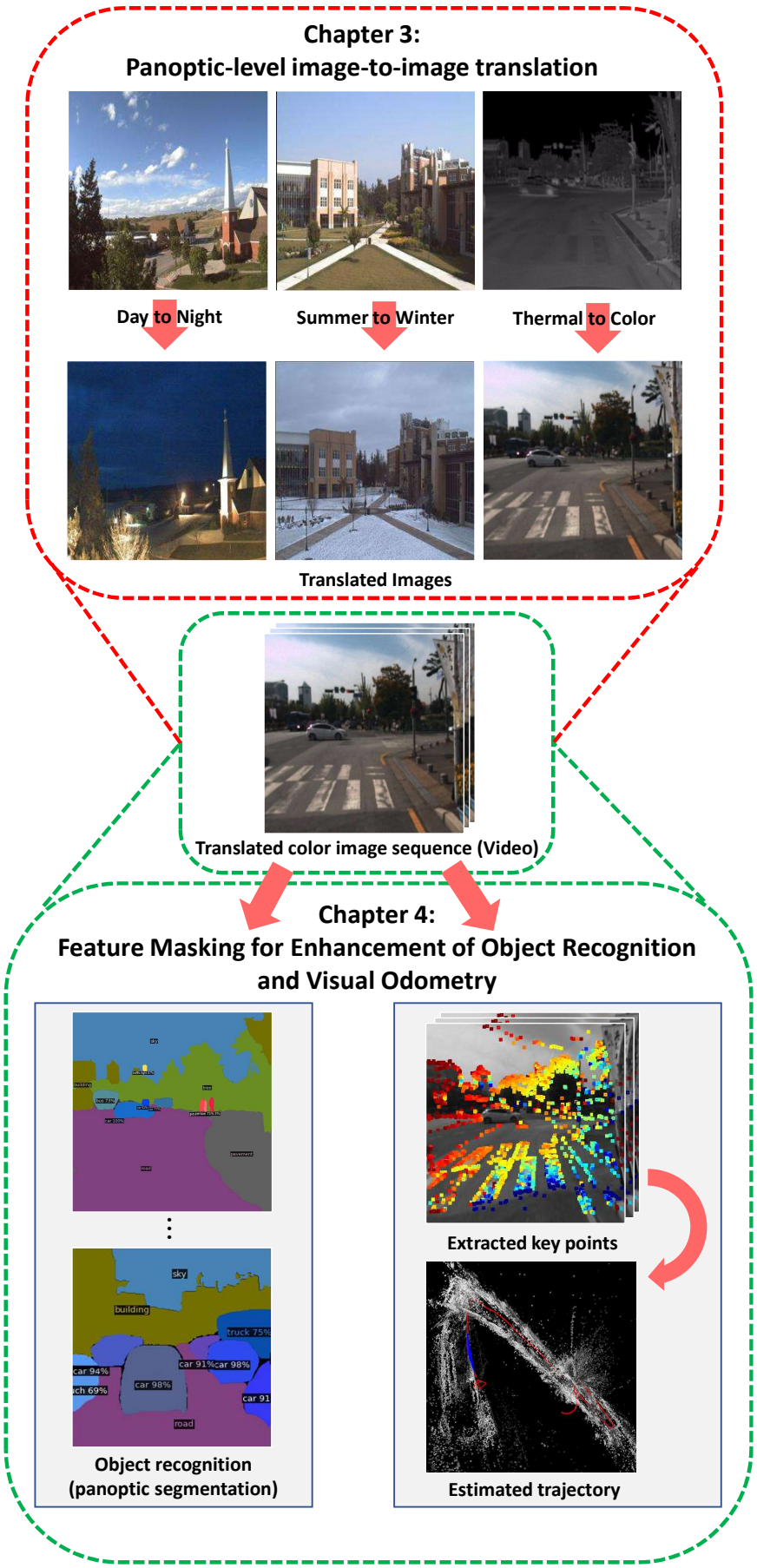


Figure 1.6. Overview of the dissertation

Chapter 2

Literature Review

This chapter presents the related works on three main topics that are important for the development process of this dissertation. There is the image-to-image translation, instance-level image-to-image translation, panoptic-level image-to-image translation, and enhancement for visual odometry (VO).

2.1 Image-to-Image Translation

Image-to-image (I2I) translation models transform the input domain image to the target domain [14], changing the style but keeping the scene content unchanged. Pix2Pix [25] 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 [29] can achieve unpaired dataset training by using cycle consistency loss.

The disentangled representation models [14], [15] utilize a combination of the content from the input domain and the style from the target domain for unsupervised learning. TICCGAN [30] adds perceptual loss [19] and total variation (TV) loss [31] to optimize networks. Pix2PixHD [26] can translate high-resolution images via a multiscale discriminator and coarse-to-fine generator. PGGAN [32] utilizes a progressive growth strategy to synthesize images from low to high resolution. AGGAN [33] and U-GAT-IT [34] can extract attention regions as structure guidance to localize important content for high-quality results. TSIT [35] uses a two-stream generative model with a feature transforma-

tion in a coarse-to-fine fashion for capturing and fusing the multiscale semantic structure information and style representation.

FDIT [36] proposes a novel frequency domain image translation framework, exploiting the frequency information to enhance the image generation process. Pang et al.[37] 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 [38] 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 [39] generator.

However, the above image-level I2I translation methods only consider the image-level information not refine the object-level features in the translation.

2.2 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 [40] and sg2im [41] can synthesize images from object scenes, and Layout2Im [42], OC-GAN [43], and LostGAN [44] 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 [45] incorporates a set of instance attributes for instance-aware I2I translation. DA-GAN [46] uses a deep attention encoder to enable the instance-level correspondences to be discovered. SCGAN [20] and SalG-GAN [47] utilize saliency-based guidance for image translation, which prioritizes salient regions and is considered an instance-level method due to the higher saliency values of foreground objects.

Shen et al.[2], Su et al.[48] and Chen et al.[49] combined an instance-level feature with an image-level feature for higher quality instance-level I2I translation. InstaFormer [50] 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.

2.3 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.[51] extracted image regions for the discriminator to improve the performance of the GANs. Huang et al.[52] provided a solution to semantically control output based on references. Dundar et al.[53] proved that panoptic-level perception makes the generated images have a higher fidelity and shows the tiny objects in more detail. Semantic segmentation [54, 55, 56, 57] can obtain a semantic perception of an image, but it cannot identify instances. Instance segmentation [58, 59, 60, 61] can obtain the object instance perception of an image, but it cannot segment the uncountable background regions. Panoptic segmentation [11, 62] 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 [28] network to extract panoptic perception (covering 'thing' and 'stuff') to combine the sampled panoptic-level style codes and image-level style codes with the correspond-

ing 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.

2.4 Enhancement for Visual Odometry (VO)

Feature-based methods [7, 8] use feature matching to estimate the camera poses via minimizing the re-projection error. Direct methods [9, 10] directly optimize the photometric error without feature descriptors.

In order to improve VO performance in challenging poorly conditions, NID-SLAM [63] replace intensities with a newly designed metric considering entropies in the frame. Alismail et al.[64] proposed a direct VO method with binary feature descriptor to enhance a poor light environment. Meanwhile, a few learning-based methods have been also proposed compared to the classical methods. Gomez et al.[65] use an LSTM-based neural network to enhance images and evaluated different methods on real-world static scenes. Multi-frame GANs [12] transformed low light images to refined images for a satisfactory SLAM performance.

Compared to above methods, we use a panoptic-level thermal-to-color image translation neural network to generate fine-grained color images from thermal images. We validate results on DSO for VO performance comparison of extracted key points and estimated trajectories.

Chapter 3

Panoptic-level image-to-image translation

3.1 Introduction

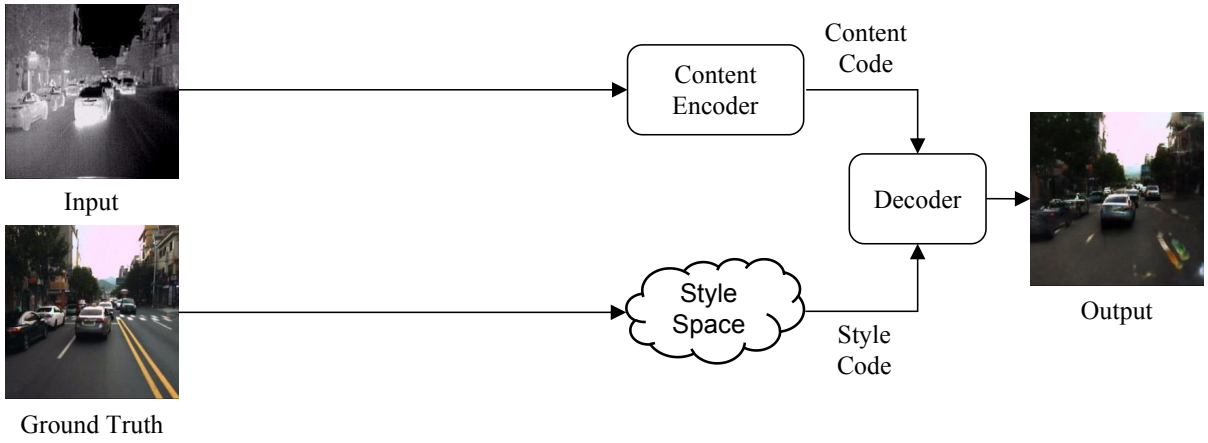
In this chapter, the author describes a pioneering and universal panoptic-level image-to-image translation network model. First, the concept and advantages of the panoptic-level model are presented via comparisons with image-level and instance-level methods. Second, the overview description contains the training manner and pipeline of the proposed panoptic-level model. Third, the details of model architecture and loss functions are shown. Moreover, the implementation details are also described in detail.

3.2 Novelty

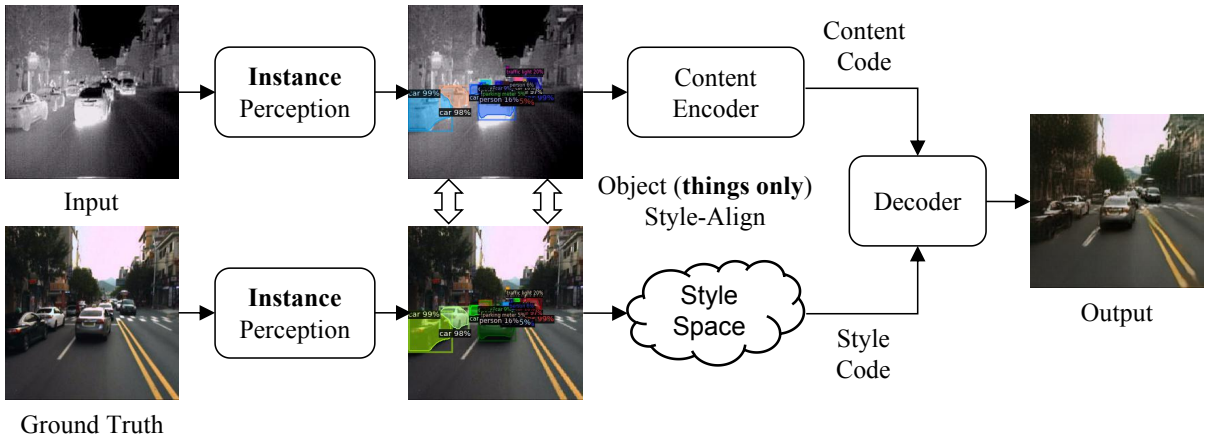
In order to illustrate the concept and advantages of the proposed panoptic-level model concisely, we compared the image-level image-to-image (I2I) translation, the object-level I2I translation, and the proposed panoptic-based I2I translation frameworks respectively.

Image-level I2I translation: This approach uses a content encoder to extract image representation from the original image directly as content code. As illustrated in Figure 3.1 (a), the style code is sampled from the style space of the ground truth image [1]. They are combined to input the decoder for the image-level I2I translation.

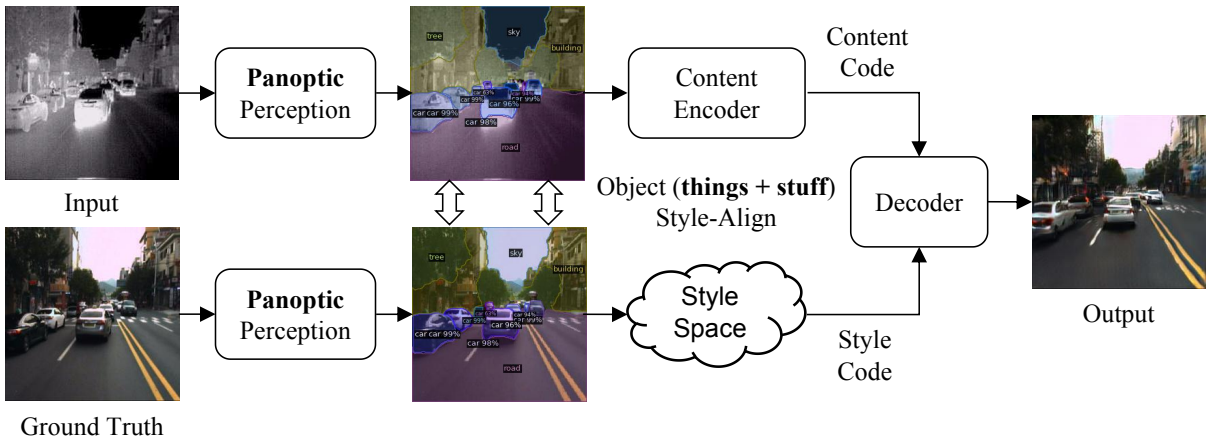
Instance-level I2I translation: This approach uses instance segmentation [27] to perceive the object information (i.e., bounding boxes, categories, and



(a) Image-level I2I Translation



(b) Instance-level I2I Translation



(c) Panoptic-based I2I Translation

Figure 3.1. Illustration of comparison for the Image-level image-to-image (I2I) translation, object-level I2I translation and our proposed panoptic-based I2I translation.

masks) in the original image. In Figure 3.1 (b), the content encoder uses it to extract object instance representations as content code, and the instance-level style code is also sampled from instance-level style space based on instance perception of ground truth image. The object instances from the input domain and ground truth domain are aligned to achieve precise object-level style transformation [2]. However, the instance perception of object-level I2I translation method only considers foreground object instances (e.g., car, person, which are countable objects termed "things" in panoptic segmentation [11]), and does not consider background semantic regions (e.g., sky, road, which are uncountable objects termed "stuff").

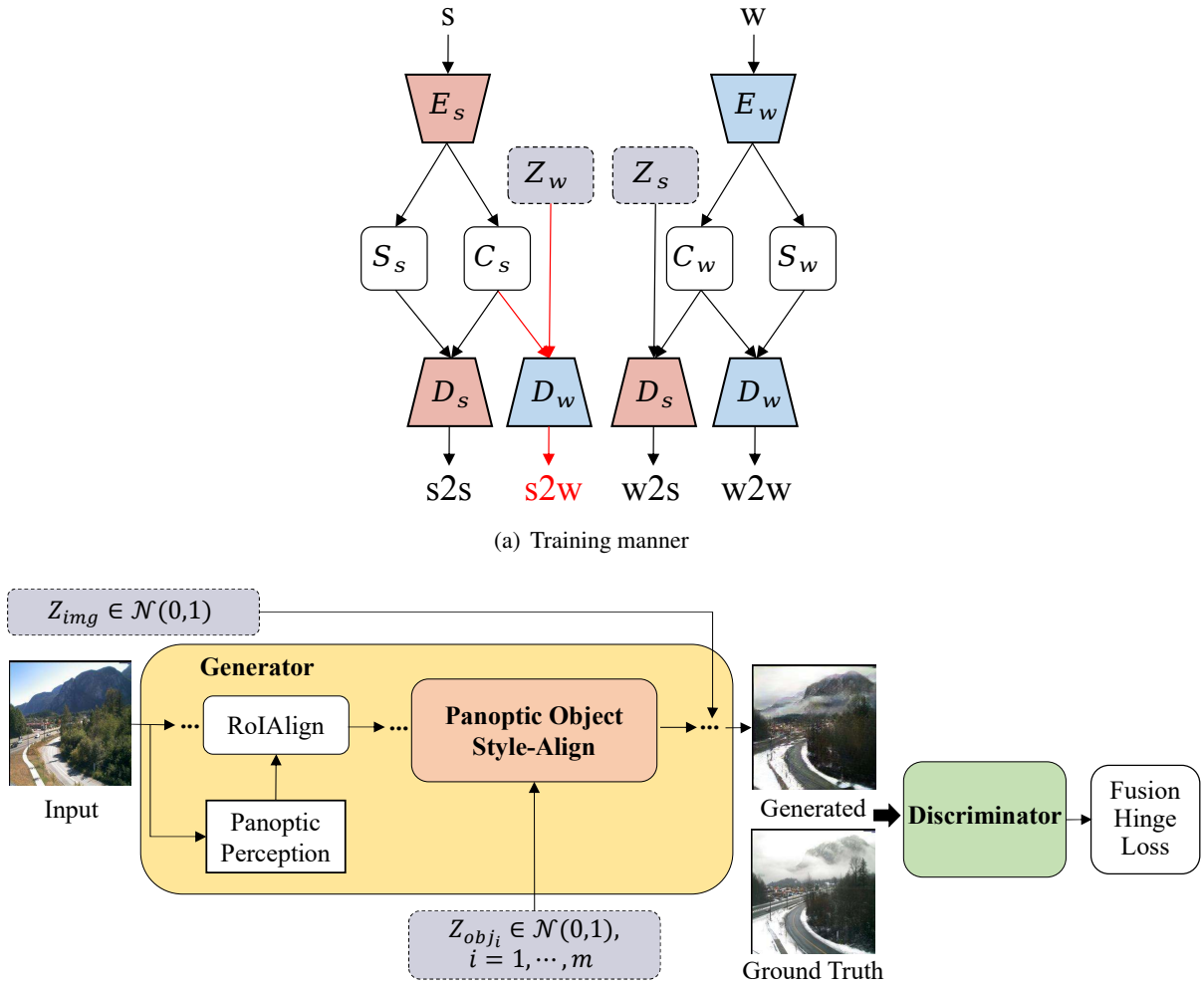
Panoptic-based I2I translation: We utilize panoptic segmentation [28] to extract panoptic perception, which considers both things and stuff. As illustrated in Figure 3.1 (c), compared with object instance-level (things only) style-align of object-level I2I translation, our method can achieve object panoptic-level (things + stuff) style-align. This demonstrates that our method has the superiority of fine object perception and more precise style transformation in the I2I translation process, which can achieve translated images in higher fidelity and tiny objects in more detail in complex scenes with multiple discrepant objects.

3.3 Overview

We provide an overview of the 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.

3.3.1 Training manner

In Fig. 3.2 (a), we use a summer image s and a winter image w from the Transient Attributes dataset [66] 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



(a) Training manner

(b) Pipeline

Figure 3.2. 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 pre-trained panoptic segmentation model to support the basic RoIAlign module and proposed module (panoptic object style-align). The sampled image-level Z_{img} and panoptic-level Z_{obj} style codes are combined for target image generation.

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 [14].

3.3.2 Pipeline

In Fig. 3.2 (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 of the generator. They are provided to the basic RoIAlign [27] module and our proposed novel module (panoptic object style-align), 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 [67].

3.4 Architecture

Our architecture, as illustrated in Fig. 3.3, is built on a generator, a discriminator and the proposed novel module (panoptic object style-align). We deploy a generative adversarial learning setting via the summer and winter domain images from the Transient Attributes dataset [66] to illustrate our architecture.

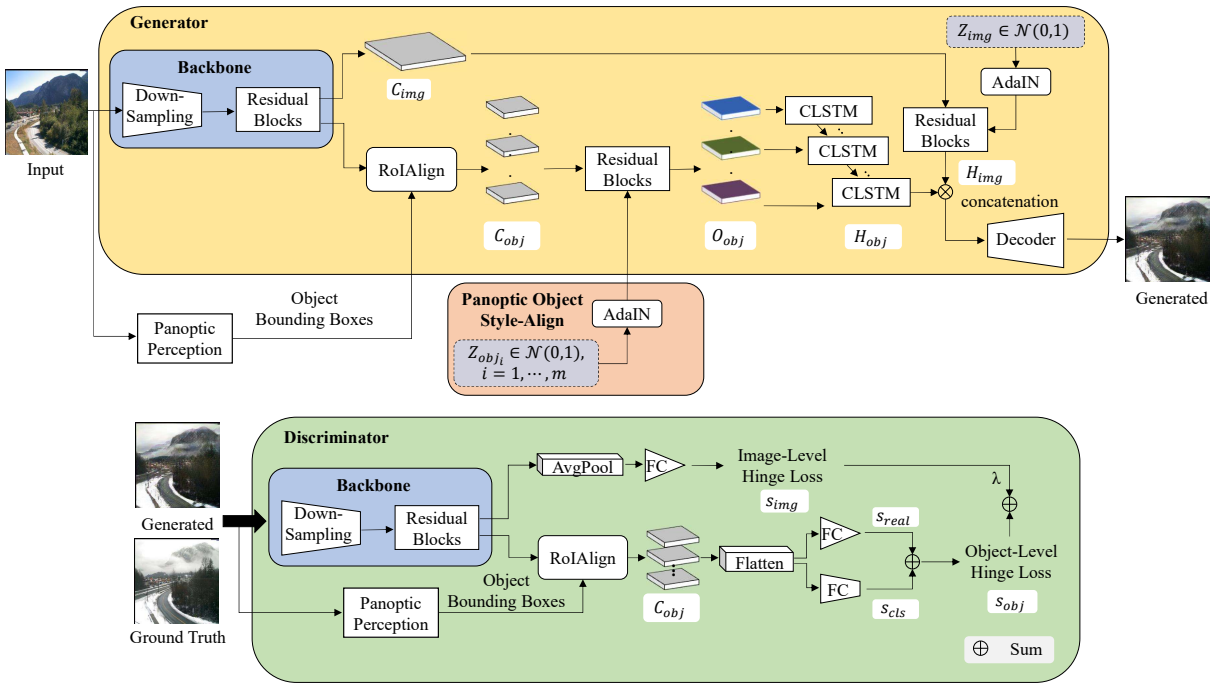


Figure 3.3. Architecture. In the generator, the summer image is extracted by the backbone consisting of down-sampling residual blocks to image-level representations, which are cropped by RoIAlign by bounding boxes from the panoptic perception of the pretrained model to the panoptic-level representations. Both representations are aligned with sampled winter style codes from the normal distribution. The results are reintegrated by cLSTM to combine with image-level representations for winter 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.

3.4.1 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 $P = \left\{(category_i, bbox_i, mask_i)_{i=1}^{m}\right\}$ 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 [11], here the 'thing' has 80 categories and the 'stuff' has 54 categories). C_{img} is cropped

by RoIAlign [27] 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}):

$$w^{'} = G(s|Z_{img}, Z_{obj}; \Theta_G) \tag{3.1}$$

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) [39] 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.

$$AdaIN(x_i, y) = y_{\gamma,i}\left(\frac{x_i - \mu(x_i)}{\sigma(x_i)}\right) + y_{\beta,i} \tag{3.2}$$

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 $O_{obj} = \{O_{obj_i}\}_{i=1}^{m}$.

$$O_{obj} = AdaIN(C_{obj}, Z_{obj}) \tag{3.3}$$

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}.

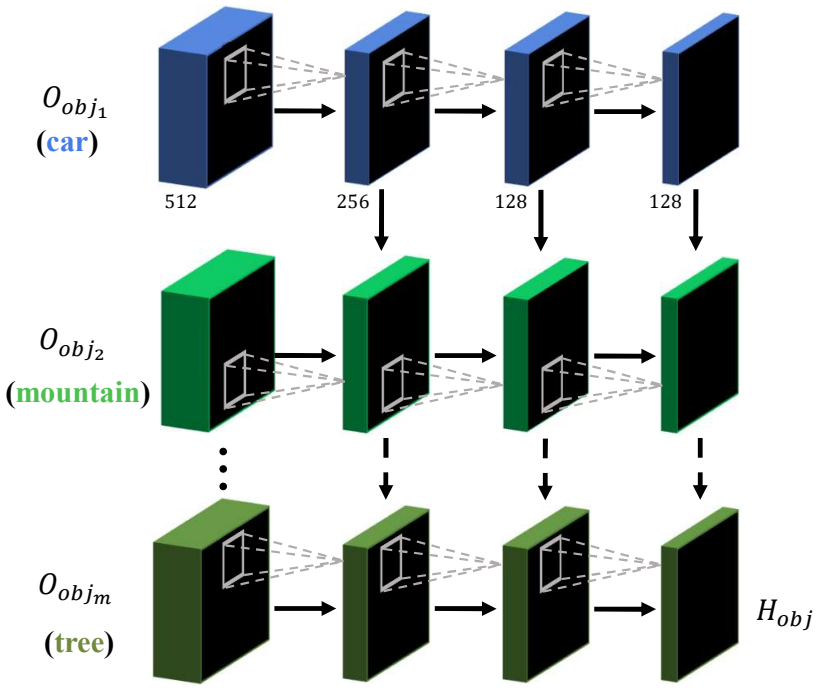


Figure 3.4. Illustration of the convolutional long short-term memory (cLSTM) component. We use three layers of cLSTM to fuse all the object feature maps into hidden feature maps H_{obj}. The numbers of channels in each layer of cLSTM are 256, 128 and 128. The residual blocks are omitted for clarity. The first of each row is the object feature maps of $O_{obj} = \{O_{obj_i}\}_{i=1}^{m}$.

As illustrated in Fig. 3.3, O_{obj} contains m object feature maps $\{O_{obj_i}\}_{i=1}^{m}$, each object feature map needs 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. 3.4, the convolutional long short-term memory (cLSTM) [68] 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 [69]. 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, $\{O_{obj_i}\}_{i=1}^{m}$, 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 $w^{'}$.

3.4.2 Discriminator

As illustrated in Fig. 3.3, 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 [27] as object content codes $C_{obj} = \left\{C_{obj_i}\right\}_{i=1}^{m}$ 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 [70, 71, 44] 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.

$$(s_{img}, s_{obj_i}, \cdots, s_{obj_m}) = D(I; \Theta_D) \tag{3.4}$$

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.

3.5 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.

3.5.1 Adversarial Loss

We utilize the image-level and object-level fusion hinge version [67] of the standard adversarial loss [72] 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.

$$l_k(I) = \begin{cases} \min(0, -1 + s_k); & \text{if } I \text{ is ground truth } w \\ \min(0, -1 - s_k); & \text{if } I \text{ is generated } w^{'} \end{cases} \tag{3.5}$$

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 [44],

$$\begin{aligned} L_{\text{adv}}(\Theta_D, \Theta_G) &= -\mathop{\mathbb{E}}_{(I)\ p_{all}(I)}[l(I)] \\ L_{\text{adv}}(\Theta_G, \Theta_D) &= -\mathop{\mathbb{E}}_{(I)\ p_{fake}(I)}[D(I; \Theta_D)] \end{aligned} \tag{3.6}$$

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.

3.5.2 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 train-

ing manner. Image reconstruction loss is essential for within-domain training and is also a very efficient pixelwise image generation optimization.

3.5.3 Perceptual Loss

We use L_{p} to alleviate the problem that the translated images are prone to producing distorted textures [26]. It is beneficial to keep the textures in a high-level space through the ReLU activation of the VGG-19 network [19],

$$L_{\mathrm{p}} = \sum_k \frac{1}{C_k H_k W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} \left\| \phi_k(w')_{i,j} - \phi_k(w)_{i,j} \right\|_1 \tag{3.7}$$

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.

3.5.4 Full Objective

The final loss function is defined as:

$$L_{\mathrm{total}} = \lambda_1 L_{\mathrm{adv}} + \lambda_2 L_1^{img} + \lambda_3 L_{\mathrm{p}} \tag{3.8}$$

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.

3.6 Implementation Details

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

3.6.1 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.

3.6.2 Training

To stabilize the training of GANs, spectral normalization [73] 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 [74] 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 [75].

3.7 Experiments

We conduct extensive experiments to evaluate our method with competing I2I translation baseline tasks to show our superiority. The evaluation of the image

quality mainly conducts a comprehensive comparison based on three aspects, realism, sharpness and diversity. Based on the specific metrics, we compare our method against several baselines, both qualitatively and quantitatively.

3.7.1 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 [14], BicycleGAN [1], TSIT [35], FDIT [36] and pSp [38] were categorized under image-level I2I translations. As SCGAN [20] uses saliency maps for salient regions perception, we regarded SCGAN as an instance-level translation. INIT [2] and InstaFormer [50] are instance-level methods, so we also implemented them as instance-level evaluations for fairer comparisons.

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 [28]. 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.

3.7.2 Datasets

We trained and evaluated our model on the Transient Attributes [66] and KAIST-MS [3] 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 [28] pretrained on the COCO-Panoptic dataset [11] 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 [3] dataset.

The unaugmented source data from our contributed dataset contained 2,026 pairs of thermal and color images based on the partial KAIST-MS [3] 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].

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

3.7.3 Evaluation Metrics

We evaluated the image quality by choosing the Human Preference (HP), Inception Score (IS) [76], Fréchet Inception Distance (FID) [77] and Diversity Score (DS) metrics instead of the Peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) [78] metrics. For images generated by the GAN learning models, the traditional PSNR and SSIM metrics deviate from human visual perception [79].

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.

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 [80],

$$IS(G) \approx exp\{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{1000} p(y=j|I_i) log\frac{p(y=j|I_i)}{\hat{p}(y=j)}\} \tag{3.9}$$

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.

Fréchet Inception Distance (FID)

FID improved the inception score (IS) [76] by incorporating statistics from real images. Similar to IS, FID also uses an Inception V3 network pretrained on ImageNet to compute the fréchet distance [81] between two Gaussian distributions fitted to the synthesized images and real images [80]. Therefore, the lower the FID is, the better a generator model is. The fréchet distance is defined by,

$$\begin{aligned} FID^2((\mu^0,\Sigma^0),(\mu^1,\Sigma^1)) = \|\mu^0-\mu^1\|_2^2 + \\ T_r(\Sigma^0+\Sigma^1-2(\Sigma^0\Sigma^1)^{1/2}) \end{aligned} \tag{3.10}$$

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.

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 [42]. We used the LPIPS metric [79] for diversity scoring, and the pretrained AlexNet [82] for feature extraction.

$$DS(I_1,I_2) = \sum_{i=1}^{n} \frac{1}{|\Lambda_i|} \sum_{p\in\Lambda_i} \left\| w_i \odot (x_1^i(p) - x_2^i(p)) \right\|_2^2 \tag{3.11}$$

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.

Table 3.1

Human Preference is 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 indicates better image quality.

Model	Human Preference ↑		
	t2c	d2n	s2w
MUNIT	2.40%	1.23%	3.36%
BicycleGAN	1.10%	3.14%	2.41%
TSIT	3.22%	10.01%	3.35%
SCGAN	1.23%	6.37%	1.59%
SCGAN+Pan	–	–	–
INIT	9.51%	2.24%	11.87%
INIT+Pan	–	–	–
FDIT	16.07%	5.93%	8.74%
pSp	11.34%	20.49%	14.78%
InstaFormer	24.68%	18.12%	23.92%
InstaFormer+Pan	–	–	–
Ours	**30.45**%	**32.47**%	**29.98**%

3.8 Result

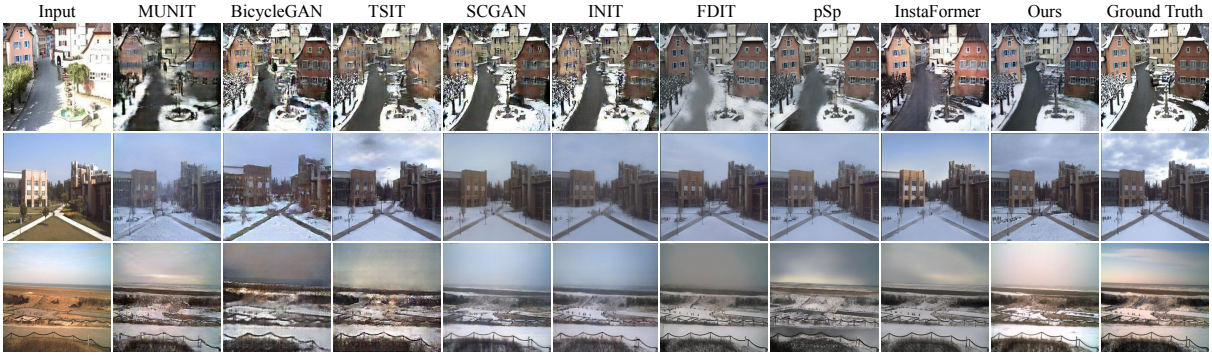
3.8.1 Qualitative Results

For image quality, the human preference results in Table 3.1 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.

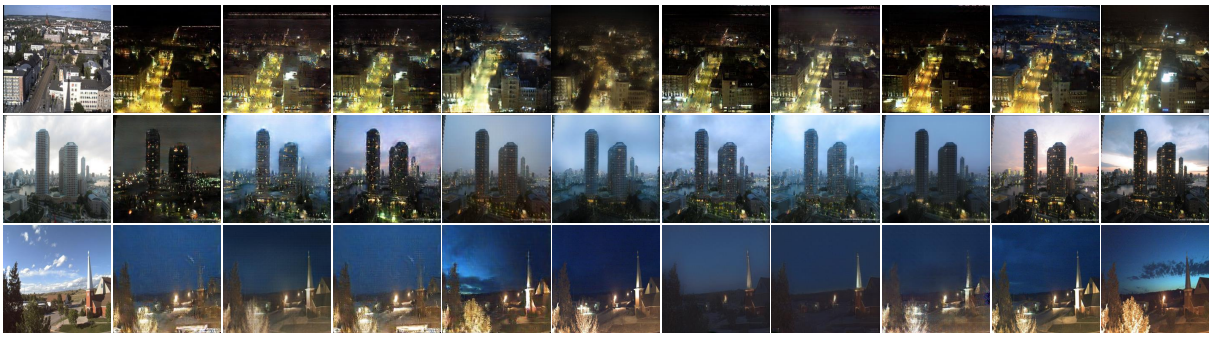
Fig. 3.5 demonstrates that our PanopticGAN could translate to higher fidelity and brightly colored images and show tiny objects in more detail. In contrast, the results from other methods were more blurry, distorted and missing small objects. In the image quality comparison of Fig. 3.5, although the 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.

3.8.2 Quantitative Results

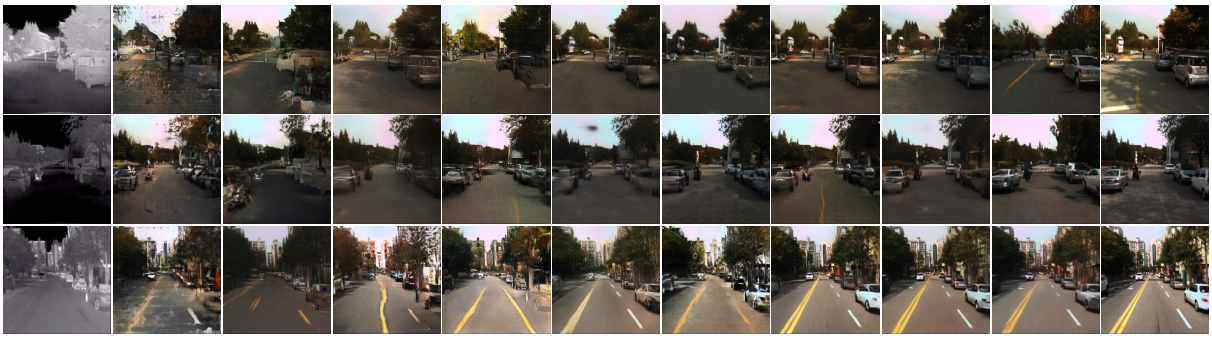
For the image quality, the inception score, fréchet inception distance and diversity score in Table 3.2 show that our approach achieved superiority in the image



(a) **Summer → Winter**



(b) **Day → Night**



(c) **Thermal → Color**

Figure 3.5. 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.

quality of the translated images compared with other approaches.

Overall, our method 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

Table 3.2

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 Inception Score, higher Diversity Score and lower Fréchet Inception Distance indicate better image quality.

Method	Inception Score ↑			Fréchet Inception Distance ↓			Diversity Score ↑		
	t2c	d2n	s2w	t2c	d2n	s2w	t2c	d2n	s2w
MUNIT	2.29	1.50	1.92	98.51	98.79	93.97	0.46	0.65	0.62
BicycleGAN	2.61	1.86	1.81	98.82	97.96	92.23	0.47	0.60	0.61
TSIT	2.64	1.78	1.96	95.38	80.86	81.32	0.43	0.67	0.64
SCGAN	2.59	1.62	1.58	96.83	92.40	86.45	0.39	0.53	0.49
SCGAN+Pan	2.62	1.68	1.67	91.21	90.67	83.20	0.42	0.57	0.53
INIT	2.70	1.22	1.84	83.26	76.73	78.90	0.37	0.65	0.57
INIT+Pan	2.75	1.53	1.86	79.34	70.82	74.67	0.45	0.66	0.61
FDIT	2.63	1.59	1.90	90.31	72.65	74.23	0.49	0.68	0.60
pSp	2.73	1.68	1.91	79.08	73.86	76.27	0.52	0.70	0.63
InstaFormer	2.80	1.89	1.95	74.71	72.34	73.52	0.51	0.69	0.64
InstaFormer+Pan	2.82	1.90	1.98	73.64	70.83	72.19	0.50	0.70	0.67
Ours	**2.85**	**1.93**	**2.01**	**72.73**	**69.40**	**71.16**	**0.54**	**0.72**	**0.69**

highly robust when a scene was invariant, especially for the objects generated on the image.

3.9 Discussion

In the development of PanopticGAN, we considered the balance between maintaining the semantic integrity of the original image and introducing the style of the target domain. One of the key elements in our design was the use of panoptic perception; this includes both foreground instances and background semantic regions. These were used as content codes and aligned with the sampled panoptic style codes. This approach not only avoided the loss of semantic information but also allowed for a rich fusion of content and style codes for each object, leading to higher-fidelity results.

As for the model collapse problem when training GANs, we set up different learning rates and optimization strategies for the generator and discriminator. Also, we added additional noise to the latent space. As for the convergence problem, we used the fusion hinge loss for training. Also, we set up the spectral normalization for the generator and discriminator, respectively.

Our method uses panoptic-level perception to avoid losing too much information in the translation. For less extreme input images (e.g., summer2winter), we can use pre-trained models from benchmarks (e.g., Panoptic-DeepLab) to replace annotation cost. Otherwise, we can get perception from paired images for training and only annotate a compact dataset for inference (e.g., thermal2color).

However, the proposed method was not without challenges. As we use the bounding boxes from panoptic perception to crop each object feature map rather than using masks, it is well known, that bounding boxes cannot precisely crop out object bountries since each object has a different shape. This, in turn, could lead to less sharp object boundaries in the translated images. Furthermore, while our PanopticGAN generally improved the fidelity of the translated semantic regions, there remained potential for unsatisfactory performance in certain object recognition or visual odometry tasks. These issues highlight the need for further refinement of our method.

Despite these challenges, PanopticGAN offered a significant advancement in image-to-image translation. The use of panoptic perception as prior knowledge improved the quality of the translated semantic regions, which contributed to the overall effectiveness of our model. The introduction of panoptic-level information has shown to be a promising direction for future developments in this field.

3.10 Conclusion

The author proposed a novel panoptic-aware image-to-image translation method (PanopticGAN). The panoptic perception is extracted to achieve a panoptic-level combination between content and style codes. The image-level combination of content and style codes is also merged for translating images with high fidelity and tiny objects in more detail in complex scenes with multiple discrepant objects. Compared to the image-level and instance-level methods, the proposed panoptic-level method refines the object semantics on both the foreground and background to avoid semantic information loss and improves the fidelity of the translated objects to further enhance task performance.

3.11 Contribution

The contributions of the work presented in this chapter were:

- 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.

Chapter 4

Feature Masking for Enhancement of Object Recognition and Visual Odometry

4.1 Introduction

In this chapter, the author describes a feature masking module that is proposed to precisely extract the representations within each object contour by masks and is integrated into the original model architecture for sharpening object boundaries and translating higher fidelity objects. This design enhances the performances of object recognition and visual odometry due to the translated higher fidelity objects of images/videos.

First, the advantages of the feature masking module are presented, and a new model architecture integrating the feature masking module is described. Second, the enhancement of object recognition and visual odometry is discussed emphatically since they are two critical components within the field of computer vision, machine learning and robotics. These components serve as the groundwork for various applications, including vision perception, autonomous navigation and interactive robotics. Moreover, extensive experiments are conducted in detail, and the experimental results are comprehensively compared to demonstrate the superiority of the proposed approach.

4.2 Feature Masking

As illustrated in Fig. 3.3, according to the original model architecture principle, the content codes and style codes of the image panoptic objects are fully aligned in the panoptic object style-align module, and then each object-level representation in the result is directly used the convolutional long short-term memory (cLSTM) component to fuse into the image-level representations, which are used to finally generate the target image.

However, there are still certain problems with this architecture, that is, the representations of objects are extracted by the RoIAlign [27] module into representations of specific sizes based on their respective bounding boxes and then operated. We know that the shape of each object is irregular, therefore there is a lot of information redundancy in object representation based on bounding boxes. Removing the redundant information in the representation of each object is our motivation, that is, extracting the true representation of the object more accurately. To address this challenge, we propose a novel module called feature masking.

4.2.1 Overview

To do this we need to show the pipeline after adding feature masking to illustrate the details. As shown in Fig. 4.1, we take the conversion of thermal images to color images as an example. Based on the original pipeline, we introduced the proposed feature masking module. We use a pre-trained panoptic segmentation network to obtain the panoptic perception of the ground truth color image paired with the input thermal image, it provides panoptic-level bounding boxes, categories, and masks. The bounding boxes are provided to RoIAlign [27] and masks are provided to the proposed feature masking module.

Firstly, image-level representation is extracted, panoptic-level style codes Z_{obj} and image-level style codes Z_{img} are sampled from the normal distribution, $Z_{obj} = \left\{Z_{obj_i}\right\}_{i=1}^{m}$ is processed by the proposed panoptic object style-align module, m is the number of objects perceived in the panoptic perception. We treat both 'thing' and 'stuff' as objects in panoptic perception. Z_{obj} and Z_{img}

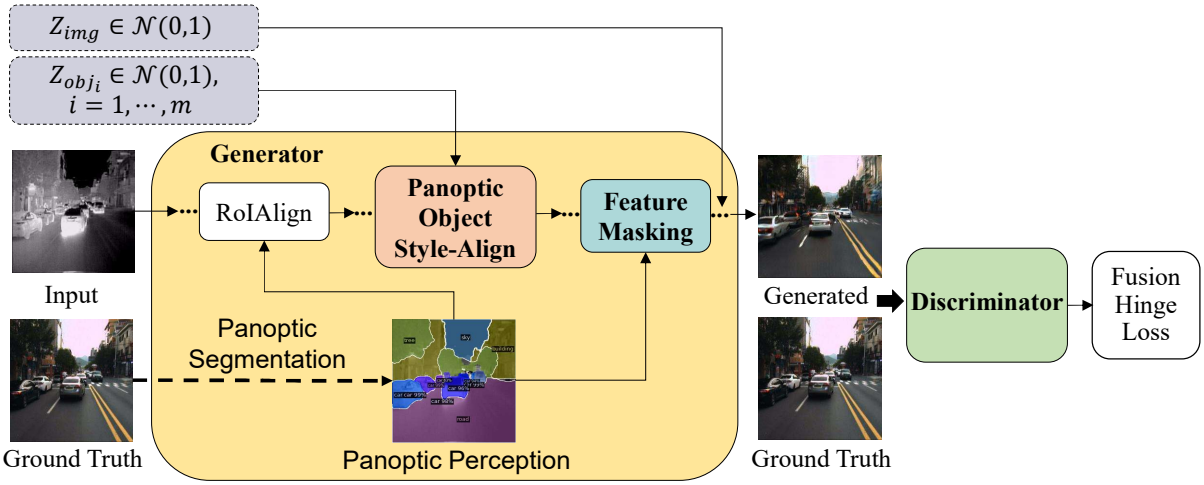


Figure 4.1. Overview. A panoptic perception is extracted from the pretrained panoptic segmentation model to support the basic RoIAlign module and proposed modules (panoptic object style-align and feature masking). The sampled image-level Z_{img} and panoptic-level Z_{obj} style codes are combined for target image generation.

will be aligned with corresponding panoptic-level and Image-level representations in the generator for panoptic-level image translation. The translated color images are fed into the discriminator, where we use fusion hinge loss consisting of image-level and object-level adversarial hinge loss terms [67] for optimization.

4.2.2 Architecture

We also take the conversion of thermal images to color images as an example to illustrate the new architecture after adding the proposed feature masking to the original model architecture. As shown in Fig. 4.2, the input is a thermal image, and its paired ground truth color image provides panoptic perception through a pre-trained panoptic segmentation model. We introduce the module details of feature masking based on the original model architecture. The object representations are processed by the panoptic object style-align module and then processed by the proposed feature masking.

For the feature masking module, as illustrated in Fig. 3.3, O_{obj} contains m object feature maps $\{O_{obj_i}\}_{i=1}^{m}$. Since the object bounding boxes $P(bbox_i)_{i=1}^{m}$

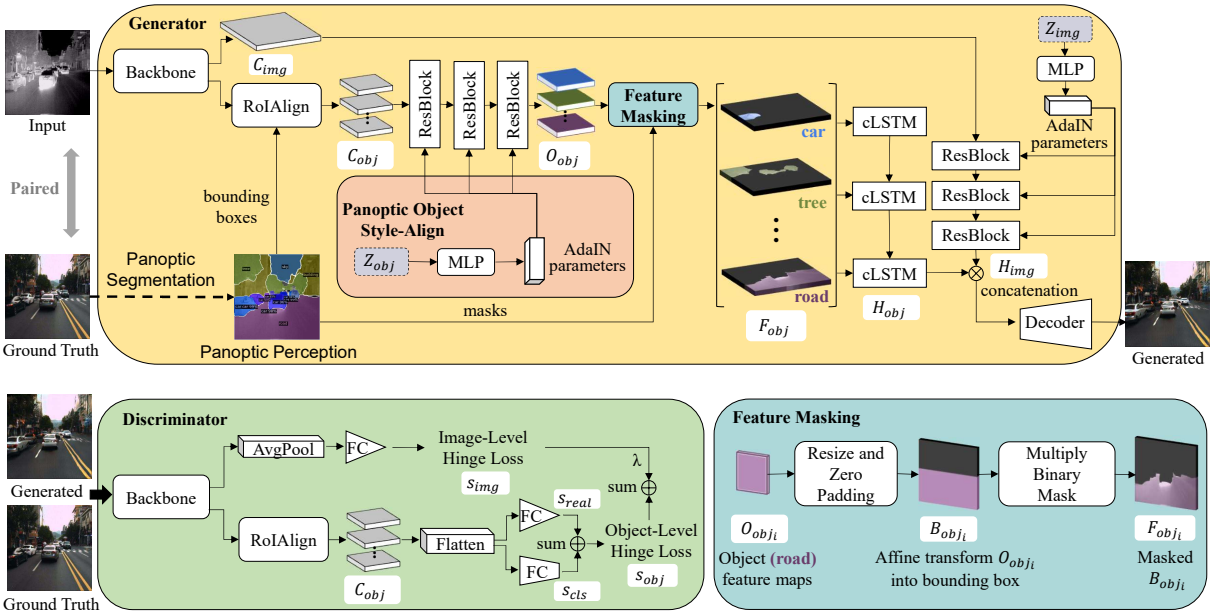


Figure 4.2. Architecture. In the generator, the thermal image is extracted by the backbone consisting of down-sampling residual blocks to image-level representations, which are cropped by RoIAlign by bounding boxes from the panoptic perception of the pretrained model to the panoptic-level representations. 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.

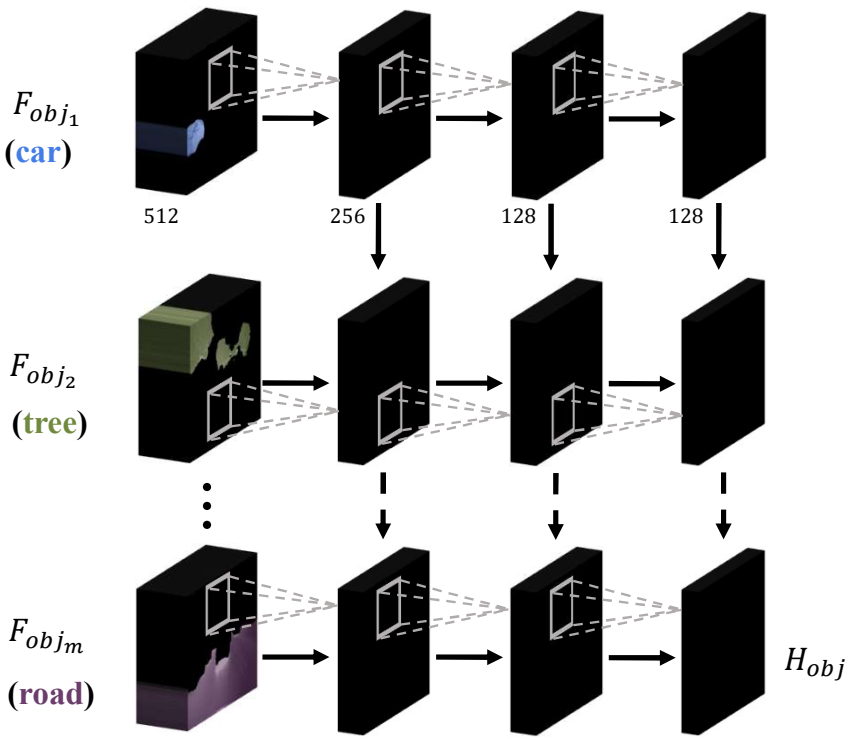


Figure 4.3. Illustration of the convolutional long short-term memory (cLSTM) component. We use three layers of cLSTM to fuse all the object feature maps into hidden feature maps H_{obj}. The numbers of channels in each layer of cLSTM are 256, 128 and 128. The residual blocks are omitted for clarity. The first of each row is the object feature maps of $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

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 [83] 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 $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

$$F_{obj} = B_{obj} \cdot M \tag{4.1}$$

After feature masking, the masked object feature maps need to be fused into a well-hidden representation to generate a realistic target color 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. 4.3, same as the original details, the new cLSTM 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 color image.

4.3 Enhancing Object Recognition

Object recognition is a significant aspect of computer vision, normally the first step towards enhancing object recognition lies in the architecture of the neural networks used. In recent years, the field has seen significant advancements in neural network designs, with deep learning models such as Convolutional Neu-

ral Networks (CNNs) and self-attention networks taking the spotlight. Also, the quality and diversity of training data are also fundamental to the success of any machine learning model. Exploring various methods of augmenting training data, including the generation of synthetic data and transformation techniques can enhance the model's ability to generalize and perform accurately in diverse settings thereby improving the performance of object recognition.

Our work involves the development and application of PanopticGAN, an innovative model based on Generative Adversarial Networks (GANs), traditionally used for image-to-image translation tasks. The PanopticGAN has been meticulously designed to excel in complex scenarios, particularly target object semantics generation. This approach has led to more realistic translated images, which significantly bolster the quality of object recognition.

We will focus on utilizing the experimental evaluation of the proposed PanopticGAN to verify the improvement in the performance of object recognition. The application of image translation learning has proven instrumental in increasing its efficiency in the object recognition task.

4.4 Enhancing Visual Odometry

Simultaneous Localization and Mapping (SLAM) is a critical technique for autonomous navigation, especially in environments where GPS signals are weak or unavailable. The integration of SLAM techniques with visual odometry can lead to more accurate and reliable navigation systems. Visual odometry, a crucial aspect of autonomous navigation systems, can be significantly improved through the integration of depth sensing.

Visual odometry systems rely heavily on accurate depth information to estimate motion and understand scene geometry. The integration of depth information from stereo vision and LiDAR technology has significantly improved the accuracy of these systems.

Also, machine learning-based optimization can improve the performance of visual odometry systems. Applying machine learning algorithms, such as reinforcement learning can optimize these systems for real-world applications. The

application of these algorithms has led to an improvement in the adaptability and performance of visual odometry systems in dynamic environments.

In our work, we utilize the sequential high-quality image frames translated by the proposed PanopticGAN to deploy the visual odometry systems. This has led to a substantial improvement in the accuracy of these systems, particularly in estimating motion and understanding scene geometry. By providing high-quality image translations through PanopticGAN, we managed to enhance the performance of SLAM-based systems significantly.

The PanopticGAN's high-quality image translations have been instrumental in this optimization process, contributing to improved visual odometry performance across various scenarios. We will focus on utilizing the experimental evaluation of the proposed PanopticGAN to verify the improvement in the performance of visual odometry.

4.5 Experiments

We conducted extensive experiments on our model that added the feature masking module for task performance enhancement, which compared with baseline methods to show our superiority. We mainly evaluate two aspects, i.e., the object recognition performance of the translated images and the visual odometry performance of the translated videos.

For the object recognition performance, we mainly perform panoptic segmentations on the translated images and compared the panoptic quality (PQ) [11] 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) [10]. 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 se-

quences and the testing sequences.

Based on the specific metrics, we compare our method against several baselines, both qualitatively and quantitatively. The comparisons include the object recognition performance on translated images and the visual odometry performance on translated videos.

4.5.1 Baselines

For the baselines, since the experimental results of TICCGAN [30] 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.

4.5.2 Evaluation Metrics

We mainly evaluated the object recognition performance of the translated images and the visual odometry performance of the translated videos.

We chose the Panoptic Quality (PQ) [11] series metrics to evaluate the object recognition performance. The absolute pose error [84] metric was used for the visual odometry performance.

Panoptic Quality (PQ)

PQ was adopted to evaluate the object recognition performance, and PQ combines segmentation quality (SQ) and recognition quality (RQ) [11],

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \tag{4.2}$$

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. $\mathrm{PQ}^{\mathrm{Th}}$,

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.

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 [84].

We ran the DSO [10] 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.

4.6 Result

4.6.1 Qualitative Results

After adding the feature masking module in the originally proposed model, 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. 4.4 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.

For the object recognition performance, we used the panoptic segmentation result from the pretrained Panoptic FPN model on the COCO-Panoptic dataset.

Figure 4.4. 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.

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 of insignificant differences for an object recognition comparison.

As illustrated in Fig. 4.5, 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 [30] to the baselines since its results showed a superior performance. Therefore, we selected four methods (i.e., TICCGAN [30], INIT [2], pSp [38] and InstaFormer [50]) with the highest performance to compare with our results. We ran the

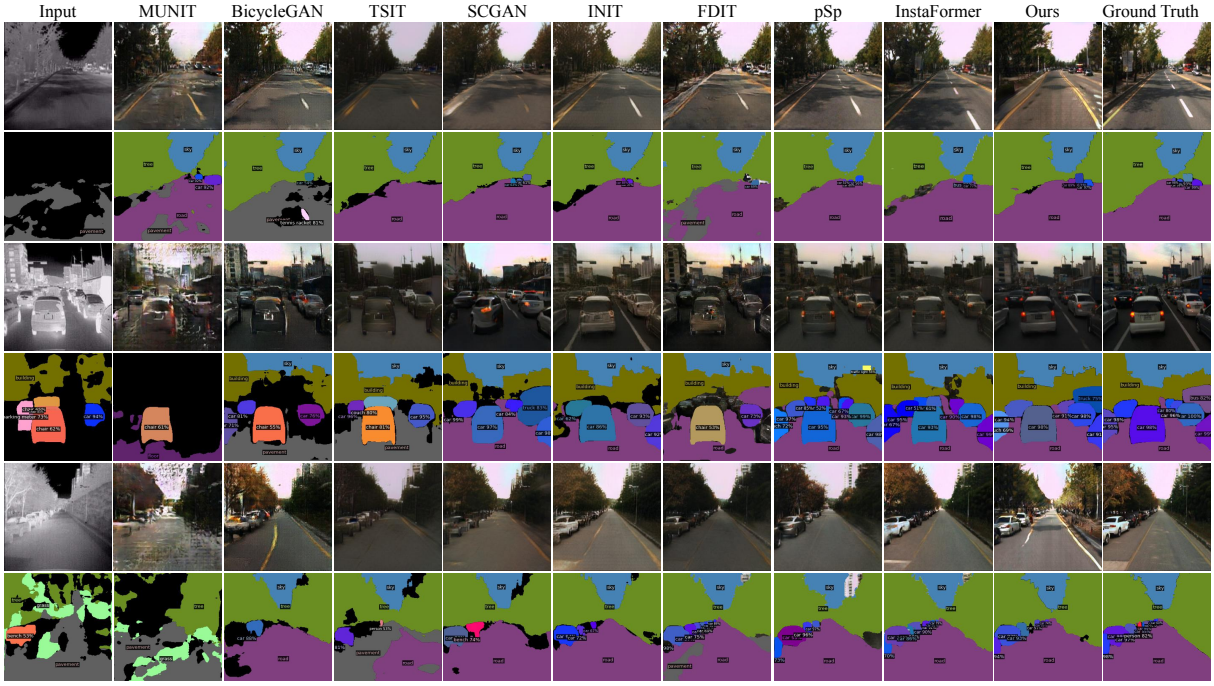


Figure 4.5. The object recognition performance of the translated images from different approaches. In each group, the upper row shows the translated images, and the lower row shows the results from the corresponding panoptic segmentation. Our method tends to achieve better results, especially for tiny objects.

DSO [10] method on both the original thermal frame sequences (video) and the corresponding translated color frame sequences through different competing approaches.

First, Fig. 4.6 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. 4.7 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-

Table 4.1

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 [14]	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 [35]	6.4	17.2	8.1	2.1	13.3	3.3	13.9	26.4	15.3
SCGAN [20]	5.6	15.2	7.4	1.7	11.8	2.6	13.6	22.5	17.4
SCGAN+Pan	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+Pan	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+Pan	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**

MS dataset [3] 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.

4.6.2 Quantitative Results

For the object recognition performance, Table 4.1 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 [10] on the trans-

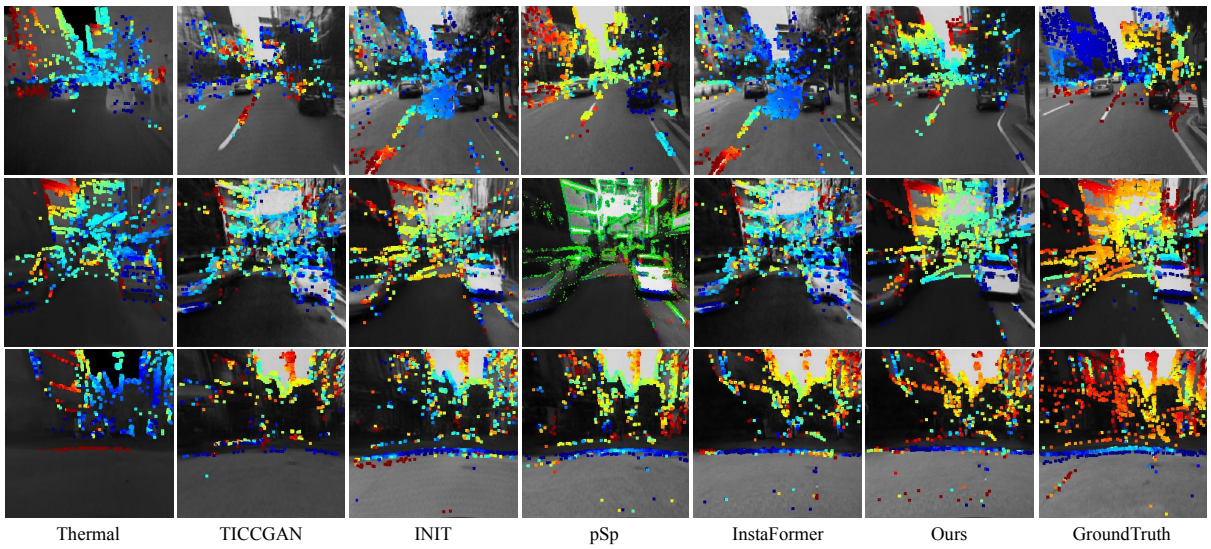


Figure 4.6. The key point comparisons of the translated frames in the visual odometry for the different approaches. The blue points indicate a large depth value; the red points indicate a small value; and the green points indicate the initialized value. Our results have more robust points and are more consistent with the ground truth.

lated videos. As expected, Table 4.2 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. 4.7. 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 (t_{rel}) and the rotational error (r_{rel}) were lower than the others, which was consistent with the estimated trajectory comparison results in Fig. 4.7. 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.

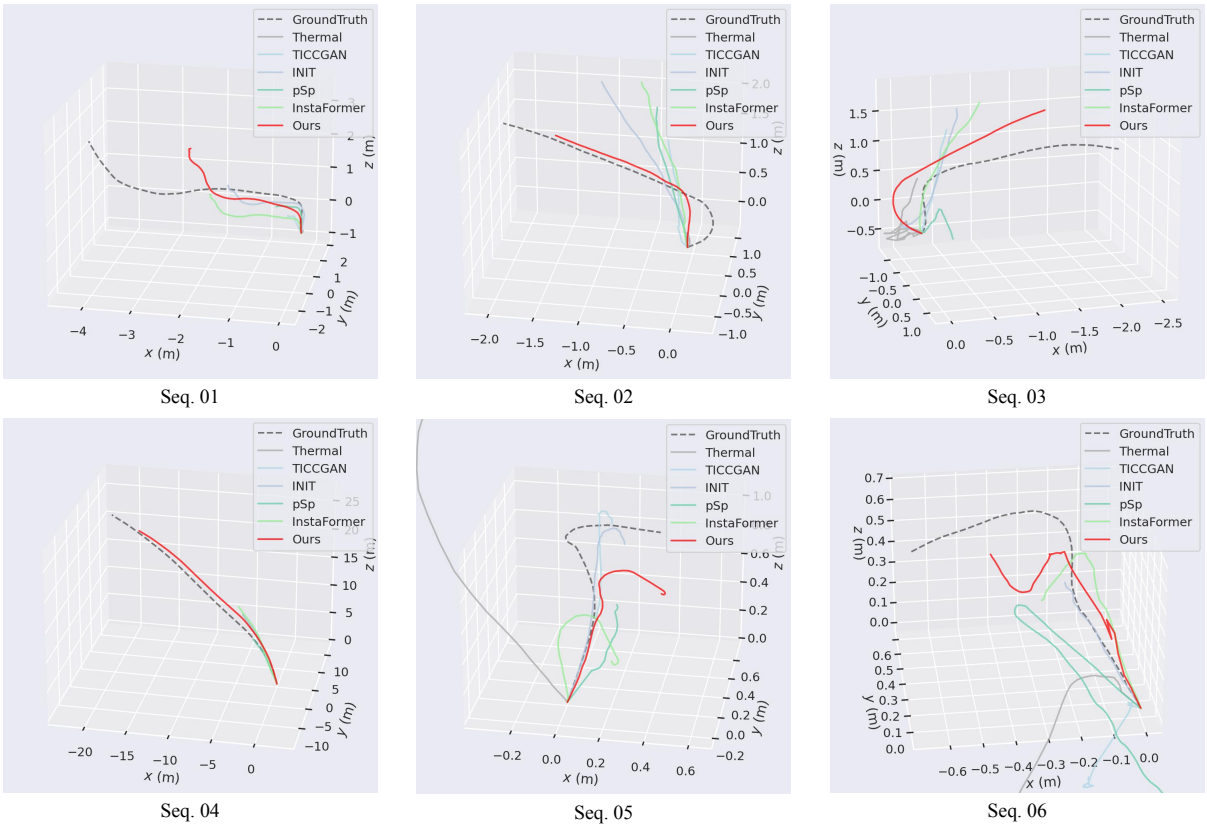


Figure 4.7. The performance comparisons of the translated subsequences (video) in the visual odometry for the different approaches, mainly for the trajectories. Our estimated trajectories tend to be more consistent with the ground truth.

4.6.3 Ablation Study

We demonstrate the necessity of all the losses and modules of our proposed model by comparing the inception score (IS) [76], fréchet inception distance (FID) and diversity score (DS) [42] for the image quality; panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ) for the object recognition performance [11]; and absolute translational error (t_{rel}) and absolute rotational error (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 [3] dataset of the thermal-to-color translation task.

As shown in Table 4.3, 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,

Table 4.2

The t_{rel} and r_{rel} metrics (lower is better) are utilized to evaluate the visual odometry performance (trajectory).

	Thermal		TICCGAN		INIT		pSp		InstaFormer		Ours	
Seq.	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	1.41	1.60	1.07	1.24	**0.89**	**1.16**
02	2.45	1.81	1.75	2.79	1.59	2.86	1.45	1.97	1.38	1.71	**1.28**	**1.49**
03	2.37	3.14	1.31	1.29	1.20	1.38	1.17	1.26	1.12	1.20	**1.03**	**1.10**
04	2.21	2.42	2.03	1.89	1.34	1.60	1.08	1.14	0.97	1.01	**0.32**	**0.45**
05	3.30	2.56	2.44	2.23	2.18	2.46	2.23	2.17	2.14	2.09	**2.05**	**1.71**
06	6.51	7.02	7.41	7.52	4.05	3.17	6.21	5.53	3.24	3.36	**2.16**	**2.34**

Table 4.3

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 (t_{rel}) and absolute rotational error (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**
t_{rel}	2.03	1.43	1.50	1.41	2.07	1.81	**1.28**
r_{rel}	2.10	1.63	1.60	1.67	1.93	1.78	**1.37**

SQ, and RQ were all significantly decreased because L_{obj} computed the category projection scores for the objects; the t_{rel}) and 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.

4.6.4 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 4.4, we present the model efficiency comparisons between our scores and the best baselines' scores. Here, we list the competing baselines, TSIT [35], INIT [2], INIT+Panoptic (INIT+Pan), pSp [38], InstaFormer [50] and InstaFormer+Panoptic (InstaFormer+Pan). Table 4.4 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 com-

plexity, and the average processing time kept it competitive.

Table 4.4

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+Pan	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+Pan	120.3	53.4	18.2	21.4	21.1
Ours	**113.6**	51.4	**17.6**	**19.0**	19.9

Figure 4.8. 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.

4.7 Discussion

As shown in Fig. 4.8, 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 [50] method.

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.

4.8 Conclusion

The author introduced a novel feature masking module into the proposed panoptic-level image translation method. It can extract the object-specific representations along the contours to sharpen the object boundaries in generated target images. Based on this, the proposed method 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.

4.9 Contribution

The contributions of the work presented in this chapter were:

- 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 trans-

lated 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 [3] 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.

Chapter 5

Conclusion

As outlined in Section 1.8, this chapter concludes the study by discussing major research findings, addressing the research questions, and highlighting contributions made to the field. The limitations of the study and potential future research directions will also be discussed.

5.1 Summary of Major Findings

The objective of this study was to advance the performance of rescue robots in incipient disaster environments by improving their visual perception using a machine learning-based thermal vision enhancement algorithm. The core idea was to transform low-resolution and blurry thermal images into high-resolution and fine-grained colorful images to improve the practicality of visual odometry and object recognition tasks.

In Chapter 3, the study introduced a pioneering and novel panoptic-aware image-to-image translation based on the generative adversarial network, i.e., PanopticGAN. It avoids a substantial loss of semantic information, and the latent space of each object has a rich fusion of content and style codes to generate higher-fidelity images. The results showed that PanopticGAN could translate low-resolution and blurry thermal images into high-resolution and finer-grained color images, significantly improving the translated image quality. The effectiveness of PanopticGAN was evaluated, and the results demonstrated superiority over other available methods.

In Chapter 4, the study highlighted a feature masking module that is proposed

to precisely extract the representations within each object contour by masks and is integrated into the original model architecture for sharpening object boundaries and translating higher fidelity objects. This design enhances the performances of object recognition and visual odometry due to the translated higher fidelity objects of images/videos. The extensive experiment results further validate the potential of PanopticGAN with feature masking as a robust solution for enhancing the performance of rescue robots in disaster scenes.

Considering the research questions and philosophy, this study aimed to enhance the perception of rescue robots in disaster environments with machine learning algorithms. The proposed PanopticGAN structure in Chapter 3 and the enhancement of object recognition and visual odometry based on the translated results by PanopticGAN with feature masking in Chapter 4 are significant steps toward this goal.

However, there were certain limitations in the study. Such as the training is supervised, which consumes expensive annotation costs and leads to a generalization problem on different datasets. Also, in extreme cases, the high diversity of panoptic objects may result in incorrect textures being generated to large semantic regions, which may result in a decrease in the overall image quality and may impact the consistency of the continuous frames to further affect the visual odometry performance.

5.2 Future Study

In future work, we aim to overcome the limitation that supervised learning leads to expensive annotation costs and generalization problems, and 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 a new unsupervised learning way to cut costs and improve generalization ability, also 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 op-

tical 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.

Bibliography

[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 bicycle consistency," in *Advances in neural information processing systems*, pp. 465–476, 2017.

[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*, pp. 3683–3692, 2019.

[3] 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*, pp. 1037–1045, 2015.

[4] S. Khattak, C. Papachristos, and K. Alexis, "Keyframe-based direct thermal–inertial odometry," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3563–3569, IEEE, 2019.

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

[6] M. A. Marnissi, H. Fradi, A. Sahbani, and N. E. B. Amara, "Thermal image enhancement using generative adversarial network for pedestrian detection," in *25th International Conference on Pattern Recognition (ICPR)*, pp. 6509–6516, IEEE, 2021.

[7] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[8] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pp. 225–234, IEEE, 2007.

[9] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*, pp. 834–849, Springer, 2014.

[10] 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.

[11] 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*, pp. 9404–9413, 2019.

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

[13] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, pp. 700–708, 2017.

[14] 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)*, pp. 172–189, 2018.

[15] 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)*, pp. 35–51, 2018.

[16] 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.

[17] 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*, pp. 701–710, 2018.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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*, pp. 155–170, Springer, 2020.

[25] 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*, pp. 1125–1134, 2017.

[26] 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*, pp. 8798–8807, 2018.

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

[28] 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*, pp. 6399–6408, 2019.

[29] 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*, pp. 2223–2232, 2017.

[30] 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.

[31] 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.

[32] 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.

[33] 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)*, pp. 1–8, IEEE, 2019.

[34] 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.

[35] 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*, pp. 206–222, Springer, 2020.

[36] 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*, pp. 13930–13940, 2021.

[37] 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.

[38] 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*, pp. 2287–2296, 2021.

[39] 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*, pp. 4401–4410, 2019.

[40] 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*, pp. 4561–4569, 2019.

[41] 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*, pp. 1219–1228, 2018.

[42] 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*, pp. 8584–8593, 2019.

[43] T. Sylvain, P. Zhang, Y. Bengio, R. D. Hjelm, and S. Sharma, "Object-centric image generation from layouts," *arXiv preprint arXiv:2003.07449*, vol. 1, no. 2, p. 4, 2020.

[44] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10531–10540, 2019.

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

[46] 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*, pp. 5657–5666, 2018.

[47] 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*, pp. 16509–16518, 2021.

[48] 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*, pp. 7968–7977, 2020.

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

[50] 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*, pp. 18321–18331, 2022.

[51] 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*, pp. 816–825, 2021.

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

[53] 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*, pp. 8070–8079, 2020.

[54] 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*, pp. 3431–3440, 2015.

[55] 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*, pp. 2881–2890, 2017.

[56] 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.

[57] 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.

[58] 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*, pp. 6154–6162, 2018.

[59] 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*, pp. 642–651, 2019.

[60] 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.

[61] 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.

[62] 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.

[63] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, "Nid-slam: Robust monocular slam using normalised information distance," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1435–1444, 2017.

[64] H. Alismail, M. Kaess, B. Browning, and S. Lucey, "Direct visual odometry in low light using binary descriptors," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 444–451, 2016.

[65] R. Gomez-Ojeda, Z. Zhang, J. Gonzalez-Jimenez, and D. Scaramuzza, "Learning-based image enhancement for visual odometry in challenging

hdr environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 805–811, IEEE, 2018.

[66] 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.

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

[68] 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*, pp. 802–810, 2015.

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

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

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

[72] 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.

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

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

[75] 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.

[76] 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.

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

[78] 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.

[79] 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*, pp. 586–595, 2018.

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

[81] 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.

[82] 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.

[83] 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*, pp. 3992–4000, 2015.

[84] 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*, pp. 3354–3361, IEEE, 2012.