Overcoming Temperature Discrepancy in Thermal image Place Recognition using Edge-guided Multimodal Image-to-Image Translation

Dong-Guw Lee[1] Hyeonjae Gil[1] Seungsang Yun[1] Jeongyun Kim[1]
Ayoung Kim[1]
[1] Department of Mechanical Engineering, Seoul National University, Republic of Korea
{donkeymouse, h.gil, seungsang, jeongyunkim, ayoungk}@snu.ac.kr

Abstract

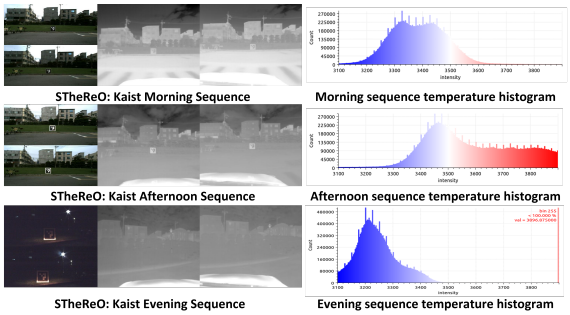
Place recognition is a crucial task in autonomous driving systems that has been actively researched. However, when it comes to thermal infrared (TIR) image-based place recognition, it has shown poor performance due to the variation in appearance caused by temperature differences throughout the day, which frequently occurs in outdoor environments. To address this, we propose a GAN-based thermal image translation model that translates thermal images captured at different times of the day into contrast-consistent and detail-preserving images, achieving time-agnostic thermal image representations. We applied this translation model as a preprocessing block to common place recognition models and achieved a top-1 accuracy of 80.69% on NetVLAD and 46% on DBoW, outperforming other baseline methods on the benchmark STheReO dataset.

1. Introduction

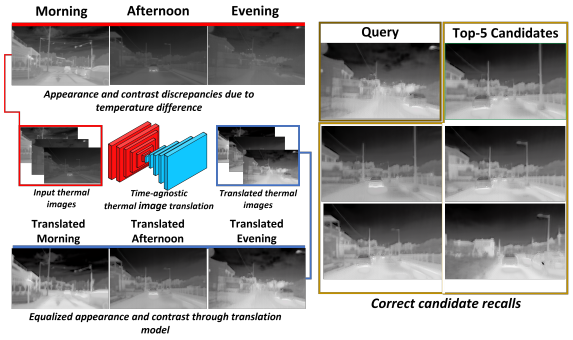
In recent years, thermal infrared images have gained popularity in computer vision applications, especially in autonomous driving related to visual perception [16] and localization [11, 20, 24]. In particular, thermal images have a distinct advantage over RGB images in that they capture emitted thermal infrared radiation from objects, as opposed to RGB images which capture reflected visible light. This unique characteristic enables thermal images to maintain their robustness in poor illumination conditions and adverse weather conditions, such as rain, snow, or storms, making them an attractive alternative for autonomous driving. Furthermore, the use of thermal images in the context of autonomous driving can provide an added level of safety in outdoor situations where visibility is compromised by illumination or adverse weather conditions.

Despite their advantages, thermal images have several limitations, such as poor resolution, low contrast, ambiguous object boundaries, and lack of color information, which limit their adaptability to methods designed for RGB images [15] More importantly, from an autonomous driving perspective, the greatest challenge in using thermal images lies in their changing visual appearance over time due to shifts in the temperature distribution of the environment. As shown in Fig. 1a, the temperature histograms of thermal images for the same scene vary depending on which time of the day it was taken from. Such appearance discrepancies caused by shifts in temperature distribution of the environment pose a critical performance degradation for thermal image-based place recognition.

In fact, classical methods for place recognition, such as DBoW, have shown limitations in thermal images due to ap-



(a) Thermal images at identical location from different time period and their respective temperature histogram. Y axis of the histogram represents the bin count and X axis represents the temperature values of each pixel expressed in 14-bit radiometric data



(b) Proposed thermal image place recognition

Figure 1. Research overview.

pearance discrepancies [21]. To overcome this limitation, previous studies have proposed solutions to minimize the appearance discrepancy between a query and a key thermal image in place recognition. Shin et. al. [21] proposed a linear transformation function to amend the intensity and contrast of thermal images, but this approach may not be applicable for longer time intervals with larger appearance differences. Saputra et al. [20] utilized a learning-based method to formulate a global descriptor using deep neural embedding for place recognition, but this approach requires paired RGB images for training and may not effectively address appearance discrepancies originating from differences in temperature in outdoor environments.

In our research, we aim to address the temporal temperature discrepancies that limit the performance of thermal image-based place recognition. Typically, we have found that thermal images taken at night have lower contrast than those taken during the day, as the environment is colder at night. In the RGB domain, previous works have used GAN-based image-to-image translation to translate night-time RGB images to daytime RGB images, thereby overcoming illumination discrepancies between two images of the same place and improving place recognition performance. Inspired by this approach, we propose a time-agnostic thermal image translation model that can minimize the distributional differences caused by temporal discrepancies in thermal images, as shown in Fig. 1b. We define "time-agnostic" in thermal imaging as the ability to translate thermal images captured at any time of day into a consistent representation, as if all images were captured at the same time of day. Essentially, our model aims to produce time-agnostic thermal image representations, where the image appearance is not affected by environmental changes.

Previous studies in RGB [1, 4] and multispectral place recognition [7] leveraged bijective and deterministic mapping based image-to-image translation models, such as CycleGAN [25]. In contrast, we employ a multi-modal image-to-image translation model to ensure the consistency of the styles of the translated thermal images, which can hardly be achieved with mapping function-based translation networks. In addition, due to poor contrasts of thermal images, key semantic attributes such as buildings and roads are hardly preserved in the translated thermal image. To prevent such loss of meaningful information, we enforce edge consistency between the original and the translated image by employing laplace of Gaussian loss. Overall, we summarize the contributions of our research as the following:

- We propose an edge-preserving thermal contrast unification translation network that yields time-agnostic thermal image representations for place recognition.
- We demonstrate the validity of our translation model in improving thermal image-based place recognition that overcomes the limitations of existing methods.
- Achieving state of the art performance in thermal image-based place recognition on the STheReO [23] benchmark dataset.

2. Related Works

2.1. Using Image translation for place recognition

Several previous studies utilized image to image translation to minimize the appearance discrepancy between the query and key image for place recognition. Many of these studies leveraged unpaired image-to-image translation leveraging cyclic consistency due to its convenience in training data acquisition. However, as CycleGAN is known for being underconstrained [19], several recent variants imposed additional loss functions [1, 4] or added contrastive learning to the baseline architecture [7, 19].

DejavuGAN [4] enforced additional edge-preservation losses to the generator and domain discriminator loss for preserving key geometric details and domain-specific characteristics. ToDayGAN [1] incorporated a three-way discriminator responsible for handling each texture, color, and gradient of the translated RGB image. However, as they both use deterministic bijective generators, the styles of the translated images in thermal images cannot be achieved. In addition, due to the low image contrast and ambiguous object boundaries in thermal images, translated images using such methods hardly preserve the key apparent attributes in the image, such as buildings and roads.

To overcome the limitations of the previous methods, we argue that a multimodal approach should be used instead, as multimodal methods disentangle images into content and style vectors, enabling the control of the style of the transferred image. MUNIT [8] first pioneered this idea, and subsequent developments such as StyleGAN [12, 13] increased the expressivity of latent vectors through learnable mapping functions. Similar to MUNIT, DRIT disentangles an image into content and style and uses a diversity loss term to encourage the generator to produce diverse outputs for the same input image. StarGANv2 [6] is a multimodal and multidomain method that utilizes learnable mapping functions in a multidomain setting by using encoded domain specific style codes. Despite the advancements in these networks to model complex attributes, our research focuses on investigating whether even a simple multimodal translation-based network would be adequate in minimizing temporal discrepancies in thermal images to enhance place recognition performance.

To preserve details in the translated images, previous studies leveraged edge-guided losses [5, 18, 22]. Luo et al. [18] leveraged multi-task learning to extract edges from images via an additional auxiliary task. However, this method requires edge ground truth labels for each image. Xu et

al. [22] proposed a CNN-based edge filter to smooth RGB images, but paired training data was required to preserve the edges in the reconstructed image. In contrast, our method achieves edge consistency without employing an additional auxiliary network or requiring paired image-edge pair data. Instead, we use Laplace of Gaussian loss to model edge consistency between two images, penalizing the translated network based on edge similarity between the input and reconstructed thermal images.

2.2. Thermal image-based place recognition

Building upon the seminal method of NetVLAD [2] for learning-based place recognition, recent studies have continued to advance RGB-image based place recognition, with models incorporating self-supervision [3], multi-scale networks [14], and advanced contrastive learning losses [7].

However, despite progress in RGB-based place recognition, studies on thermal image-based place recognition are still in their infancy, with many recent studies on SLAM [11, 21] still relying on DBoW. Shin et. al. [21] proposed a method for estimating vehicle states using sparse LIDAR points projected on thermal images, with long-term drift correction achieved using a global descriptor based on ORB features from 8-bit rescaled thermal images. To account for time-dependent differences in image intensity, an affine illumination transformation model was used, but this approach may not work well for thermal images taken at night due to low contrast and different temperature distribution.

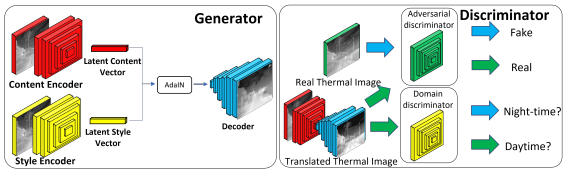
TI-SLAM [20] used a learning-based place recognition method in which the thermal images are encoded into a 128-dimension global descriptor using a CNN-based encoder and trained using triplet loss. However, its training datasets are limited to indoor and tunnel scenes with little illumination change across different time periods, which makes it difficult to generalize this method to outdoor environments where day and night thermal images have little similarity.

Most studies have conducted experiments on thermal place recognition in indoor settings, where the appearance discrepancy is minimal between a query and a key image. When tested in outdoor settings, the difficulties arising from temperature differences between two thermal images have been acknowledged, but discrepancies at longer time intervals have not been addressed. We hypothesize that by minimizing the appearance discrepancy between two thermal images using GAN-based translation networks, we could ameliorate the performance of thermal image-based place recognition for outdoor thermal images.

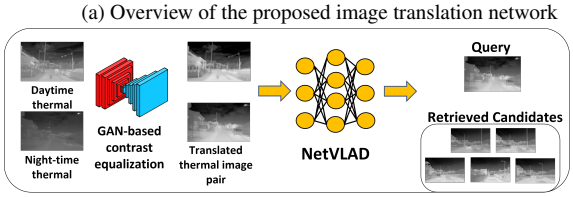
3. Time-agnostic Thermal Image Translation

3.1. Overview

Fig. 2 illustrates the proposed method. Our main idea is logical and simple. We use the thermal image translation model as a preprocessing unit to unify the intensities of a pair of thermal images. Given a pair of thermal image each of which was taken from different time but at the same location, we translate each thermal image with identical style vectors to yield pseudo-thermal images. As these translated pseudo thermal images ideally have similar distribution then the original ones, appearance wise, they should look the same. Afterwards, like shown in Fig. 2b, the two images are used as the input to a learning-based place recognition network, namely NetVLAD [2]. Our network design choices and details for thermal image-based place recognition model will be explained in the following subsection.



(a) Overview of the proposed image translation network



(b) Overview of the thermal image-based place recognition pipeline/

Figure 2. Overview of the proposed image translation network and the thermal image-based place recognition pipeline

3.2. Setting the domains

Before introducing our network architecture, it is essential to define the two domains for translating low-contrast thermal images into relatively higher-contrast thermal images. As shown in Fig. 1a, which illustrates the average histogram of all thermal images for each sequence of the STheREO dataset, the histogram of thermal images captured at night is located at the lower end of the spectrum, while the histogram of thermal images captured during daytime (morning and afternoon) is not only at a wider range but also has many components residing at the higher end of the spectrum, demonstrating higher image contrast. Therefore, we separated the dataset into the two domains in the following way: night-time thermal images, which have lower contrast, and day-time thermal images, which have higher contrast.

3.3. Network Architecture

We enhanced the MUNIT [8] baseline model by adding an edge-enhancement loss function and a domain discriminator loss to improve the appearance of translated thermal images. As shown in Fig. 2a, our network comprises shared content encoders, individual time-specific style encoders, and a single decoder. The shared content encoder

extracts common geometric features from input thermal images, while the individual style encoders capture specific thermal image characteristics, such as heat intensity and contrasts, and convert them into separate style latent vectors for each domain.

3.4. Loss functions

To train our network, we have employed loss functions commonly used in multimodal image-to-image translation networks.

Image reconstruction Loss: This loss is used to train the image translation generator via training its ability to reconstruct given input image.

$$\mathcal{L}_{recon}^{x_A} = \mathbb{E}[||G_A(E_A^c(x_A), E_A^s(x_A)) - x_A||_1] \quad (1)$$

Content and Style Reconstruction Loss: This loss enforces constraints on each content and style encoder by requiring them to reconstruct the original image from the encoded content and style vectors.

$$\mathcal{L}_{recon}^{c_A} = \mathbb{E}[||E_B^c(G_B(c_A, s_B)) - c_A||_1] \quad (2)$$

$$\mathcal{L}_{recon}^{s_B} = \mathbb{E}[||E_B^s(G_B(c_A, s_B)) - s_B||_1] \quad (3)$$

In addition, we also included in additional Laplace of Gaussian loss for edge preservation and domain discriminator loss to enhance the representation ability of the night and daytime thermal images.

Laplace of Gaussian (LoG) Loss: The LoG loss measures the similarity between the Laplacian features of the original and reconstructed images, penalizing the model for generating images with dissimilar second-order gradients at the edges. To extract the Laplacian features, 3×3 Laplacian filters are applied to each of the three image channels, followed by global-average pooling. This process is represented in (4).

$$\mathcal{L}_{Lap} = \mathbb{E}[||L(x_B) - L(x_{B,recon})||_1] \quad (4)$$

$$L(x_{DTIR}) = \frac{1}{3}(L(x^1) + L(x^2) + L(x^3))$$

Domain discriminator loss: The role of domain discriminator is to classify whether given two images are from the same domain. For this instance, given translated daytime thermal images and real daytime thermal images, the role of the discriminator is to classify whether both images were taken at the same time period. The domain discriminator is denoted as shown in Equation (5).

$$\mathcal{L}_{Domain} = \mathbb{E}[-logD_{domain}(c_x|x)] \quad (5)$$

where D_{Domain} is the domain discriminator and the $c_x x$ is the input image from specific domain c_x.

3.4.1 Overall training objective

Equation (6) denotes the overall training objective of the network. All encoders, decoders, and discriminators are jointly trained and then optimized.

$$\begin{aligned}\mathcal{L}_{\mathcal{G}} = \; & \mathcal{L}_{\mathcal{GAN}} + \lambda_{x_{recon}}\mathcal{L}_{recon}^{x} + \lambda_c\mathcal{L}_{recon}^{cNTIR} + \lambda s_{TIR}\mathcal{L}_{recon}^{sTIR} \\ & + \lambda_{Lap}\mathcal{L}_{Lap} + \lambda_{Domain}\mathcal{L}_{Domain}\end{aligned} \quad (6)$$

4. Experimental Results

From the results, we aim to observe two factors: the ability of the generative model to preserve key details in the original image in the translated image, and the ability to maintain consistent appearance between query and image candidates, particularly when they are taken at different times of the day. To quantitatively evaluate the former, we use the Average Precision Canny Edge (APCE) [17], and for the latter, we use place recognition performance. Information on the dataset and implementation details are included in the Appendix.

4.1. Thermal image translation

For comparison, we selected CycleGAN [25] and ToDayGAN [1] as baseline models since they also utilize image-to-image translation to improve place recognition performance. Due to the absence of aligned and paired night-time to daytime thermal image datasets, and the fact that our GAN translation model is unpaired, quantitative evaluation metrics such as SSIM, PSNR, and MSE that require ground truth labels cannot be directly applied to evaluate our image translation model. Instead, we used the APCE, which computes the average edge similarity between the extracted Canny edges of the original min-max normalized 8-bit thermal image and the translated thermal images at varying Canny edge thresholds. This metric evaluates the edge-guided reconstruction ability, which we believe is crucial in our proposed approach.

Table 1. Place Recognition Results on NetVLAD and DBoW

Model	NetVLAD			DBoW
Metric	Top-1	Top-5	Top-10	Top-1
Min-max	0.6803	0.8734	0.9421	0.26
CycleGAN	0.5150	0.8069	0.9013	0.06
ToDayGAN	0.5494	0.8112	0.8841	0.25
Ours (asymmetric)	0.7276	0.8951	0.9418	-
Ours (Symmetric	**0.8069**	**0.9506**	**0.9871**	0.46

The results of our image translation model are illustrated in Fig. 3. In thermal image-based place recognition, it is crucial to maintain consistent appearance between query and image candidates, especially when they are taken at different times of day and there are significant temperature discrepancies. From a qualitative evaluation, our proposed

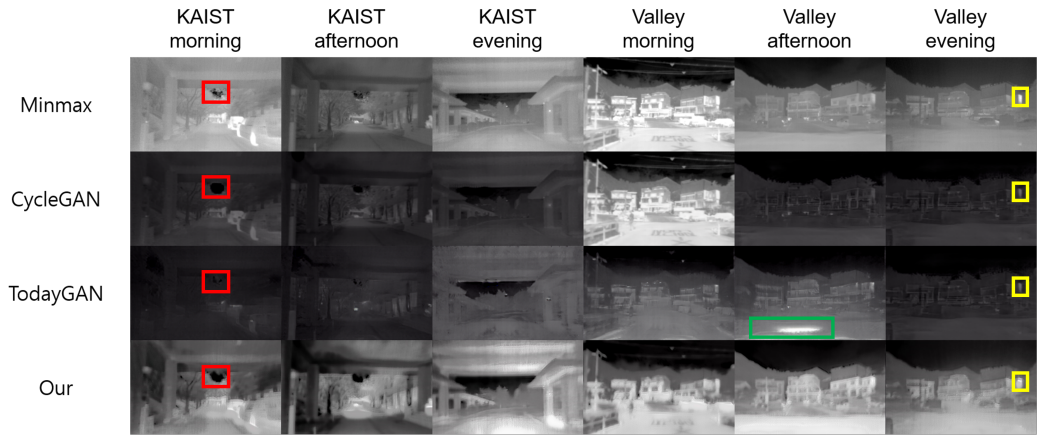


Figure 3. Image translation results on images taken from KAIST and Valley sequence of the STheReO dataset. Qualitatively, translation results using our method achieve consistent image contrasts and preserve key details (red and yellow box) in the translated image.

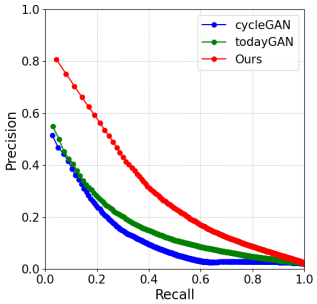


Figure 4. Precision-Recall curve of all methods. Our method showed the highest area under the curves.

translation model has achieved better inter-frame intensity consistency compared to both conventional Automatic Gain Control (AGC)-based (min-max nomralization) methods and other GAN-based methods. The scene contrast differences between daytime and nighttime are less severe in our translated images. AGC-based images exhibited drastic contrast changes between images taken at different times of the day, whereas CycleGAN translations showed less appearance variation, but with poorer contrast and inconsistency, particularly in the valley images. ToDayGAN alleviated some inter-frame consistency problems but produced artifacts and ambiguous details in the final image (see green box in Fig. 3).

On the other hand, our proposed method not only achieved inter-frame consistency and preserved key details presented in the original thermal images, but it also provided a wider and more diverse contrast range than other GAN-based methods. Specifically, the disentanglement of image enables us to maintain inter-frame intensity consistency of images that were captured at different times of the day. Furthermore, our method successfully preserves important features in the original AGC-based images, as highlighted by the red and yellow boxes. Moreover, in terms of quantitative evaluation via APCE, our proposed model demonstrated the highest overall APCE of **0.3643** for the valley sequence, when compared to CycleGAN (**0.1363**) and ToDayGAN (**0.1000**). Thus, these APCE results demonstrate that our proposed translation model is the optimal choice at preserving the key details in thermal images.

4.2. Thermal image-based Place Recognition

We evaluated the performance of our translation model for place recognition using the STheReO dataset, which includes image sequences acquired from three locations, KAIST, SNU, and Valley. We used top-1, top-5, and top-10 recall rates, which measure the ratio of successful retrievals to the total query size, as the quantitative evaluation metric. The precision-recall curve in Fig. 4 and Table. 1 show the thermal image-based place recognition recall rate performance evaluated on the STheReO dataset.

Our proposed translation method outperformed other baseline methods across all evaluation criteria. Specifically, our symmetric method (See Section 4.3.1) achieved the highest top-1 recall rate of 80.69%, compared to the conventional baseline of 68.03%. In contrast, the other baseline methods, ToDayGAN and CycleGAN, had lower recall rates of 54.94% and 51.50%, respectively. Additionally, our method had the largest area under the precision-recall curve in Fig. 4, indicating the highest precision-recall trade-off among the GAN-based models. We believe that the improvement in performance was attributed to the alleviation of appearance discrepancies between thermal images captured at different time periods, which was mainly achieved through the image contrast equalization and edge preservation capabilities of our proposed GAN model.

4.3. Ablation study

4.3.1 Symmetric vs asymmetric image translation

Our proposed image translation model was trained to equalize the appearance of both night-time and daytime ther-

mal images by minimizing the temperature-based discrepancies between them. However, given that daytime thermal images already have relatively higher contrast than night-time images, the necessity for translating both input images may be questioned. To address this, we introduce a variation to our place recognition pipeline by only translating night-time thermal images. This model is referred to as the asymmetric translation model, as opposed to the symmetric translation model that translates both query and key thermal images into joint representations. Table 1 clearly shows that the performance of the symmetric translation-based model is superior to that of the asymmetric method. One possible explanation for this result is that the symmetric model induces fewer appearance discrepancies between the translated images compared to the asymmetric model. This finding reinforces the importance of maintaining inter-frame consistency in thermal-based place recognition.

4.3.2 Performance on DBoW

In addition to evaluating our translation model with deep learning-based place recognition methods, we also tested its performance with DBoW. As indicated in the Table. 1, using our translation method led to a 46% improvement in the top-1 recall, compared to the top-1 accuracy of 26% achieved by min-max-based methods that are commonly used in thermal image-based odometry. This finding suggests that our translation model can enhance the quality of thermal images in general, rather than being exclusively designed for deep learning-based place recognition tasks. Therefore, when considering real-time performance or inference speed, our translation model, along with DBoW, can still be utilized to improve place recognition performance in thermal images.

4.3.3 Use of different latent styles

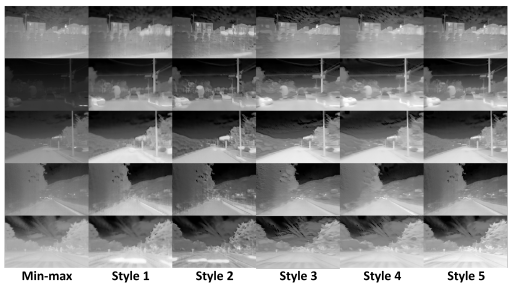


Figure 5. Image translation results when using different styles

The style code used in our proposed method affects the translated images and, consequently, the place recognition performance. To examine the effect of using different style codes on place recognition performance, we randomly sampled five latent styles, translated the thermal images, and evaluated them on NetVLAD and DBoW. Table. 2 and Fig. 5 demonstrate the quantitative and qualitative evaluations respectively. The results highlight the importance of selecting the correct style code for achieving high place recognition accuracy, as there is a notable performance gap between different styles. For example, selecting style 1 can yield a top-1 accuracy of 80.69%, while selecting style 4 can result in a top-1 accuracy as low as 60.94%. The same trend can be applied when using DBoW. From a qualitative perspective, the highest performing images retained the edges of the min-max image, indicating that edge-preservation and high contrast in the translated images are also crucial factors to consider.

Table 2. Place recognition results on the proposed method using different style latents.

Model	NetVLAD			DBoW
Metric	Top-1	Top-5	Top-10	Top-1
Style 1	**0.8069**	**0.9506**	**0.9871**	0.42
Style 2	0.6867	0.9142	0.9592	0.16
Style 3	0.7082	0.9356	0.9785	**0.46**
Style 4	0.6094	0.8755	0.9464	0.35
Style 5	0.7639	0.9335	0.9807	0.15

5. Conclusion

In conclusion, our proposed approach has demonstrated promising results in improving place recognition performance in thermal images by mitigating appearance discrepancies caused by differences in temperature. Our findings suggest a direction for future research in this field, even without the utilization of state-of-the-art techniques or advanced methods. However, more work is needed to be done to further enhance the performance of our method, and the implementation of an adaptive style selection scheme for deployment consideration is also necessary. Although we have shown improvement in performance without state-of-the-art GAN-based image translation and place recognition techniques, we believe that further progress can be made with the use of better methods. Finally, we plan to explore the integration of contrastive learning-based methods to the image translation network, but the current lack of studies on stabilizing contrastive learning in thermal images poses a challenge.

Acknowledgments This work was jointly supported by the Interdisciplinary Research Initiatives Program by College of Engineering (or College of Natural Sciences) and College of Medicine, Seoul National University (2023)and the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2023-00250727) through the Institute of Construction and Environmental Engineering at Seoul National University.

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019. 2, 4

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 3, 8

[3] Chao Chen, Xinhao Liu, Xuchu Xu, Yiming Li, Li Ding, Ruoyu Wang, and Chen Feng. Self-supervised visual place recognition by mining temporal and feature neighborhoods. *arXiv preprint arXiv:2208.09315*, 2022. 3

[4] Younggun Cho, Jinyong Jeong, Youngsik Shin, and Ayoung Kim. Dejavugan: Multi-temporal image translation toward long-term robot autonomy. In *Proc. ICRA Workshop*, pages 1–4, 2018. 2

[5] Younggun Cho, Ramavtar Malav, Gaurav Pandey, and Ayoung Kim. Dehazegan: underwater haze image restoration using unpaired image-to-image translation. *IFAC-PapersOnLine*, 52(21):82–85, 2019. 2

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2

[7] Daechan Han, YuJin Hwang, Namil Kim, and Yukyung Choi. Multispectral domain invariant image for retrieval-based place recognition. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9271–9277. IEEE, 2020. 2, 3

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2, 3

[9] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 8

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 8

[11] Jiajun Jiang, Xingxin Chen, Weichen Dai, Zelin Gao, and Yu Zhang. Thermal-inertial slam for the environments with challenging illumination. *IEEE Robotics and Automation Letters*, 7(4):8767–8774, 2022. 1, 3

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[14] Ahmad Khaliq, Michael Milford, and Sourav Garg. Multires-netvlad: Augmenting place recognition training with low-resolution imagery. *IEEE Robotics and Automation Letters*, 7(2):3882–3889, 2022. 3

[15] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. *arXiv preprint arXiv:2301.12689*, 2023. 1

[16] Dong-Guw Lee, Kyu-Seob Song, Young-Hoon Nho, Ayoung Kim, and Dong-Soo Kwon. Sequential thermal image-based adult and baby detection robust to thermal residual heat marks. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13120–13127. IEEE, 2022. 1

[17] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15808–15823, 2022. 4

[18] Yanmei Luo, Dong Nie, Bo Zhan, Zhiang Li, Xi Wu, Jiliu Zhou, Yan Wang, and Dinggang Shen. Edge-preserving mri image synthesis via adversarial network with iterative multi-scale fusion. *Neurocomputing*, 452:63–77, 2021. 2

[19] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 2

[20] Muhamad Risqi U Saputra, Chris Xiaoxuan Lu, Pedro Porto B de Gusmao, Bing Wang, Andrew Markham, and Niki Trigoni. Graph-based thermal–inertial slam with probabilistic neural networks. *IEEE Transactions on Robotics*, 38(3):1875–1893, 2021. 1, 2, 3

[21] Young-Sik Shin and Ayoung Kim. Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum. *IEEE Robotics and Automation Letters*, 4(3):2918–2925, 2019. 2, 3

[22] Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. Deep edge-aware filters. In *International Conference on Machine Learning*, pages 1669–1678. PMLR, 2015. 2, 3

[23] Seungsang Yun, Minwoo Jung, Jeongyun Kim, Sangwoo Jung, Younghun Cho, Myung-Hwan Jeon, Giseop Kim, and Ayoung Kim. Sthereo: Stereo thermal dataset for research in odometry and mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3857–3864. IEEE, 2022. 2, 8

[24] Baoding Zhou, Longming Pan, Qing Li, Gang Liu, Aiwu Xiong, and Qingquan Li. Darkloc: Attention-based indoor localization method for dark environments using thermal images. In *2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–6. IEEE, 2022. 1

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE*

international conference on computer vision, pages 2223–2232, 2017. 2, 4

6. Appendix A: Implementation Details

6.1. Thermal image translation

6.1.1 Dataset

We used night-time and day-time thermal images from two publicly available benchmark datasets, namely KAIST [9] and STheReO [23]. To avoid redundancy in the training images, the dataset was downsampled by selecting every tenth frame. In total, the dataset consisted of 8533 night-time and 8717 day-time thermal images.

6.1.2 Implementation Details

We used a PatchGAN [10] discriminator that consists of four 4×4 convolution. The generator uses a common encoder-decoder architecture that consists of content encoder, style encoder, and decoder. The content encoder consists of a single 7×7 convolution, four 4x4 convolution followed by four residual blocks. The style encoder comprises of a single 7×7 convolution, four 4x4 convolution followed by a global average pooling operation and fully connected layer. Lastly, the decoder consists of four residual blocks, two upsampling block, and a single 7×7 convolution at the end.

For training, We used input image with resolution of 640×400. For the network hyperparameters, we used Adam optimizer with a learning rate of 0.0001, weight decay of 0.5 and 0.99 as β_1 and β_2 respectively. The network was trained for batch size of 1 for 60,000 iterations.

As for the training object parameters, We optimized our image translation model by carefully selecting and fine-tuning the weighting coefficients for the loss functions, including $\lambda_{x_{recon}}, \lambda_c, \lambda_s, \lambda_{Lap}$, and λ_{Domain}, with values of 20, 10, 10, 20, and 5, respectively. We found that a high weighting coefficient for the discriminator resulted in convergence failure and mode collapse, while equal weighting for the generator-related losses led to poor contrast in the translated thermal images. Based on empirical analysis of hyperparameter tuning, we determined that higher loss weightings should be applied to the image reconstruction loss to preserve thermal image characteristics and to the Laplace of Gaussian loss to preserve edge characteristics.

6.1.3 Dataset

Table. 3 outlines the number of images included in the dataset. For each location, sequences were captured at three different timestamps: morning, afternoon, and evening. The dataset also provides ground truth annotations for place recognition derived from RTK GNSS ground truth data for each image.

6.1.4 Implementation Details

Following conventions of NetVLAD [2], we used KAIST and Valley sequences for training and validation, respectively, selecting images from the evening sequence as query and images from the daytime (morning and afternoon) sequences as database. To eliminate redundancy in the training and validation data, we selected one image from each sequence at a 2-meter interval. Overall, we used KAIST sequences (1,700 query images and 3,450 database images) for training and Valley sequences (466 query images and 944 database images) for validation.

We trained four separate NetVLAD networks on each minmax AGC images and their corresponding translated images using different image-to-image translation methods (CycleGAN, ToDAY-GAN, and our proposed model). The images were cropped to 640x400 and a batch size of 16 was used, with SGD optimizer and a learning rate of 0.0001 and weight decay of 0.5. To prevent overfitting, we adopted early stopping when there was no accuracy improvement in the place recognition evaluation for two or more epochs.

Table 3. STHeReO dataset overview

Training: KAIST / Validation: Valley			
Sequences	Morning	Afternoon	Evening
KAIST	1,716	1,730	1,700
SNU	1,795	1,811	1,844
Valley	475	465	466