Correspondence between SWIR and MWIR Images Using Augmentation and Preprocessing for Registration

Takeru Inoue†, Michiya Kibe†, and Ryusuke Miyamoto‡
†*Dept. of Computer Science, Graduate School of Science and Technology Meiji University*
‡*Dept. of Computer Science, School of Science and Technology Meiji University*
1-1-1 Higashimita, Tama-ku, Kawasaki, Japan
{takeru, kibe, miya}@cs.meiji.ac.jp

***Abstract*—Multispectral sensors are used to ensure visibility in various applications. However, when multiple sensors are used for capturing images, a misalignment may occur between the images taken by each sensor unless special care is taken. To correct such misalignments, image registration based on feature matching is conducted. However, the features captured by each sensor differ, thereby complicating the registration process. In this study, we develop an approach to overcome these challenges and to improve the registration accuracy between short-wave infrared and mid-wave infrared (SWIR and MWIR, respectively) images. First, we compare and validate SiLK, a detector-based feature matching method, and LoFTR, a detector-free feature matching method. The results clearly demonstrate the superior accuracy of LoFTR. Moreover, SWIR and MWIR images exhibit a characteristic color inversion according to Kirchhoff's law. Therefore, by inverting the color of a SWIR image and aligning the color tone between image pairs, we can improve the matching accuracy. Furthermore, by diversifying the color tones of the training data through augmentation, we can handle the domain gap between SWIR and MWIR images, thereby further enhancing the matching accuracy.**

***Index Terms*—image registration, feature matching, infrared dataset, domain gap**

I. INTRODUCTION

Effective imaging at nighttime or under other low-light conditions is essential in various applications. It generally requires specialized sensing devices instead of widely used visible-light cameras. Night vision devices that are capable of capturing images under these conditions have evolved significantly over time. Early devices were active night vision devices that relied on the reflection of emitted infrared (IR) light for imaging. Subsequent advancements leveraged improvements in imaging and amplification elements to develop low-light night vision devices that amplified scarce visible light to enable imaging. These devices, often embodied in personal wearable goggles, are widely used today [1]. However, a notable shortcoming of these devices is their dependence on a certain level of environmental light, without which their performance significantly deteriorates. As such, there is a growing interest in passive night vision devices utilizing IR sensors [2].

Passive night vision devices utilizing IR sensors can capture images in near-infrared (NIR), mid-infrared, and far-infrared wavelengths and are used in a variety of applications. For instance, NIR sensors are commonly found in commercial portable cameras, and far-infrared cameras are used in nighttime advanced driver-assistance systems. While using an IR sensor to target a single wavelength band can be effective for specific applications such as automation by a computer, it poses problems in terms of visual perception when the generated images are presented to humans. Therefore, research is currently being conducted to improve image perception by integrating images obtained from sensors covering multiple wavelength bands [3], [4].

To create an image with high visual perception from images obtained from cameras of multiple wavelength bands, it is necessary to properly integrate the information. However, unless a complex optical mechanism is incorporated, images captured from different viewpoints using different lenses must be properly integrated. Differences in lens characteristics can be compensated by performing calibration in advance. Differences in viewpoints, especially when different IR cameras are fixed in relative arrangements, can also be effectively addressed through precalibration. However, when a dispersed arrangement of individual cameras is required, a method capable of accommodating this is necessary. Therefore, this study is conducted to develop a method for finding corresponding points to accurately perform geometric transformations for image integration from images obtained from IR cameras of different wavelength bands.

Studies have actively investigated matching feature points between multiple visible images. These include early research on SIFT [5] and SURF [6] and the introduction of refined handcrafted features like KAZE [7] and AKAZE [8]. Recently, studies have widely investigated deep learning techniques that achieve dramatic improvements in accuracy, such as SuperPoint [9], SuperGlue [10], SiLK [11], and LoFTR [12]. While these methods have been confirmed to be somewhat effective when targeting visible images, when applying them to images obtained from cameras of different wavelength bands, problems such as the difference in the characteristics of images due to the wavelength bands must be considered. When using learning-based methods, such differences in characteristics can be recognized as a domain gap because feature points having different features should be classified as the same. To improve the matching accuracy, such a domain gap must be overcome even though dealing with domain gaps is generally difficult [13], [14].

In this study, we focus on two high-accuracy methods, SiLK, a detector-based feature matching method, and

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

LoFTR, a detector-free feature matching method, to establish correspondences between feature points across images of different wavelength bands. As SiLK and LoFTR are both learning-based methods, how learning is performed can significantly affect the accuracy. Therefore, in this study, we aim to improve the feature matching accuracy by adopting an approach similar to domain generalization methods to overcome domain gaps. Specifically, we implement data augmentation by varying the hue of the training data and perform preprocessing of the inference data to reduce domain gaps. For evaluating the accuracy, we use images obtained from multiple helicopter-mounted IR cameras of different wavelength bands. The errors in the converted images for image registration, one of the applications of image matching, was adopted as an evaluation criterion.

The rest of this paper is organized as follows. Section II introduces existing datasets composed of IR images and feature matching methods for visible-light images. Section III proposes preprocessing for the inference process and data augmentation for the training process for image matching after analyzing the matching methods used for visible-light images. Section IV presents the evaluation of the proposed scheme using IR images proposed in [15]. Finally, Section V presents the conclusions of this study.

II. Related Work

A. IR Image Dataset

IR image datasets may contain images captured at different wavelengths [16]–[18]. The KAIST Multispectral Pedestrian Dataset [16] contains pedestrian detection data captured using uncooled IR and color cameras. The TNO Image Fusion Dataset [18] contains night images captured using visible light, NIR, and long-wavelength IR. Such multispectral data might differ slightly in terms of the sensor positioning, angles, and other shooting parameters for each wavelength band. Therefore, unless complex optical mechanisms are incorporated, image registration is required. For example, the Multimodal Stereo Dataset [19]–[21] contains scenes in which image alignment is not achieved. In the RGB-NIR dataset [22], the relative transformations of the images are known to be estimated using SIFT [5] and RANSAC [23] for achieving image registration of RGB and IR images. As many IR image datasets consist of pairs of RGB and IR images, many studies have performed matching between visible and IR images; however, few have performed matching between IR images of different wavelength bands.

One study collected data from short-wave infrared and mid-wave infrared (SWIR and MWIR, respectively) images taken at night to create highly visible images [15]. Pilot evaluations indicated that inverting the intensity of SWIR images and registering them with MWIR images improved the visibility. However, SWIR and MWIR images had to be registered by a manual process, making it unsuitable for real-time applications. Therefore, the present study aims to perform automatic registration by using feature matching of data.

B. Feature Matching

Feature matching is one of the techniques used in image registration. Registration is conducted as follows. First, feature matching is performed, and then, outlier detection is performed using outlier detection methods such as RANSAC [23]. Next, the transformation matrix is estimated, and this result is used to perform a homography transformation to ultimately achieve image alignment. Therefore, to enhance the accuracy of registration, it is essential to use an appropriate feature matching method. Two main approaches are used for feature matching: detector-based methods and detector-free methods.

1) Detector-based Feature Matching: Detector-based methods are traditional techniques that first detect feature points and then match them using methods such as mutual nearest neighbor or SuperGlue [10]. Before the advent of deep learning, hand-crafted local features exhibited good performance. The main methods are SIFT [5], ORB [24], and SURF [6]. In recent years, deep-learning-based methods such as LIFT [25], MagicPoint [26], and SuperPoint [9] have become the mainstream owing to their high accuracy. At the time of writing, the most accurate detector-based method is SiLK [11]; it employs a simple architecture, implicitly learns keypoints, and is highly effective. However, detector-based methods are known to fail in difficult cases involving changes in illumination, repetitive structures, or low-texture scenarios.

2) Detector-free Feature Matching: To address the shortcomings of detector-based methods, detector-free methods have been proposed. These methods directly look for similarities between images to be matched. Therefore, they are robust even in repetitive patterns or textureless scenes. Furthermore, some of these methods perform coarse matching before fine matching to reduce the computational load. LoFTR is one of the most representative of these methods, and improved methods have been developed. For instance, topicFM [27] encodes more high-dimensional contextual information than LoFTR. MatchFormer [28] further enhances and optimizes LoFTR's feature extraction.

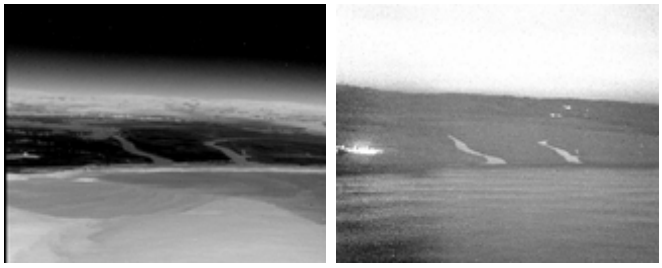
III. Correspondence between SWIR and MWIR Images for Registration

A. Overview

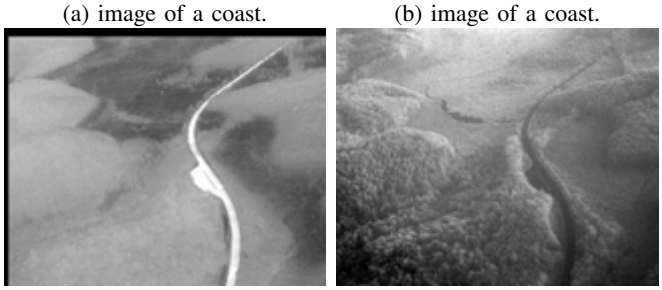
In this study, we attempt to establish an appropriate correspondence between IR images captured using different wavelengths for calibration over several cameras that may move independently. However, SWIR and MWIR images exhibit different features, as shown in Figure 1, because while an SWIR image is a reflection image similar to a visible image, an MWIR image is a thermal image that detects the IR radiation emitted by objects themselves. Therefore, SWIR and MWIR images capture different features, resulting in significantly different visual appearances even when capturing the same scene. These different features can lead to a decrease in the precision of feature matching. Therefore, this study proposes a method to reduce the accuracy loss due to these differences.

B. Comparative Analysis Feature Matching

First, we seek to determine whether detector-based or detector-free feature matching methods are more effective for our task. Therefore, we conducted comparative experiments using SiLK, a state-of-the-art detector-based method, and LoFTR, a detector-free method. We chose these two



(a) image of a coast. (b) image of a coast.



(c) image of a river. (d) image of a river.

Fig. 1: Examples of MWIR and SWIR images. The Left and right columns show MWIR and SWIR images, respectively.

TABLE I: SiLK and LoFTR achieves equivalent accuracy [11] when using HPatches [29] as evaluation data.

	Hom. Est. Acc.		Hom. Est. AUC		MMA	
	$\epsilon =$ 1	$\epsilon =$ 3	$\epsilon =$ 1	$\epsilon =$ 3	$\epsilon =$ 1	$\epsilon =$ 3
LoFTR (MegaDepth)	**0.65**	**0.87**	0.37	0.65	**0.64**	**0.91**
SiLK (top-10k)	0.62	**0.87**	**0.4**	**0.66**	0.59	0.71

methods because they demonstrated equivalent registration performance, as shown in Table I. However, detector-based methods like SiLK are known to exhibit reduced accuracy when the features of the two images are different. Therefore, we examine how much this factor influences feature matching between SWIR and MWIR images.

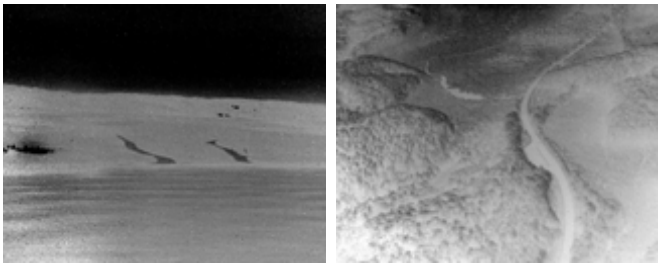
C. Preprocessing to Reduce Domain Gap

Additionally, we carried out preprocessing on pairs of SWIR and MWIR images to make them easier to match and to thereby improve the matching accuracy. In particular, we focused on the reversed hues of these image pairs. This phenomenon becomes particularly prominent at night in the absence of external light owing to Kirchhoff's law represented by the following equation:

$$r = 1 - e, \tag{1}$$

where r and e represent reflectance and emissivity, respectively. These differences in hues exacerbate the domain gap and are thought to greatly contribute to the degradation of matching accuracy.

Therefore, before performing matching, we invert the intensity of SWIR images to make it closer to that of MWIR image, thereby reducing the domain gap. Further, because short-wave IR has less light, SWIR images tend to be more prone to noise. This could make it difficult to capture texture features; therefore, we use a Gaussian filter to remove noise and improve the matching accuracy. Figure 2 shows the result of applying the inverting intensity and denoising to SWIR images.



(a) image of a coast. (b) image of a river.

Fig. 2: Results of preprocessing by inverting intensity of and denoising SWIR images.

Fig. 3: Images before and after performing data augmentation. Left and right columns show original and augmented images, respectively.

D. Data Augmentation for Model Robustness

Next, we performed data augmentation on our training data. By augmenting the data, the model learns to perform feature matching on pairs of images with different properties. This helps ensure that the model can adapt to a diverse range of data. We believe that this will improve the matching accuracy between images from different domains, such as SWIR and MWIR images. In particular, the data will be augmented to replicate different features in SW and MW. We used the following data augmentation methods:

- Randomly changing the brightness and contrast of an image
- Inverting the intensity with 50% probability
- Applying Gaussian blur with 50% probability

The brightness and contrast are changed because the brightness and contrast between SWIR and MWIR IR images are different. The inverting intensity was adopted to correspond to the phenomenon of hue inversion between SWIR and MWIR images. The Gaussian blur was adopted as MWIR images tend to be more blurry compared to SWIR images. Examples of the data augmentation results are shown in Figure 3.

IV. Experiments

A. Implementation Details

The models used for feature matching are as follows:

- SiLK (top10-k)
  See Table I. It was trained on COCO [30]. Further details can be found in the Default Setup of [11].
- LoFTR (MegaDepth)
  Outdoor model with dual-softmax that was trained on MegaDepth [31]. Details are provided in [12].
- LoFTR (Augmented)
  Model in which training data of LoFTR (MegaDepth) has been replaced with augmentation data.

B. Evaluation Dataset

The evaluation dataset employed videos captured during a flight experiment with a helicopter equipped with SWIR and MWIR sensors [15]. We utilized videos capturing scenes of forests, urban areas, coasts, rivers, and ships. Each video has a resolution of 640×512 and total frame count of 904. Additionally, the frame rates of the videos captured using the SWIR and MW IR sensors are identical, and temporal calibration has been performed.

C. Evaluation Protocol

We conduct registration using feature matching on corresponding frames. Registration begins with feature matching using the model described in Section IV.A. By using the obtained correspondence, the affine matrix between SWIR and MWIR images is estimated using OpenCV with RANSAC. Finally, image registration was performed using the estimated matrix. To evaluate the feature matching results, the registration result from the previous frame is used as the ground truth, because it is difficult to create a ground truth for feature matching itself. The average pixel error of the four corners between the registration result of the previous and current frames is calculated. After the computation of errors, the success rate of registration was obtained for three thresholds: 10, 30, and 50 pixels.

Generally, registration using feature matching should be conducted through homography transformation. However, in this study, the images included in the dataset have only position aberration owing to the positional relationship of the cameras used for acquiring SWIR and MWIR images. Therefore, affine transformation was adopted.

D. Result

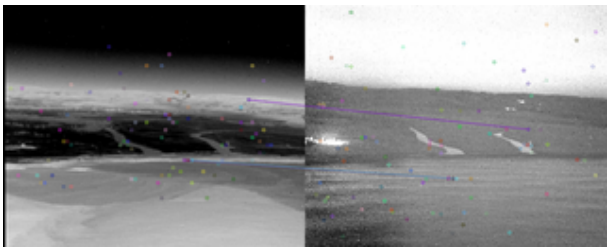
1) Comparative Analysis Feature Matching: Table II shows a comparison between SiLK and LoFTR. This table indicates that, on average, SiLK is only able to achieve less than one match per frame, resulting in a registration success rate of 0%. Therefore, SiLK is not suitable for tasks with domain gaps, such as the one in this study. These results corroborate the common belief that detector-based methods fail in scenarios with domain gaps. By contrast, LoFTR can successfully perform registration in several frames. This suggests that detector-free methods may have some level of resilience to domain gaps. Figures 4 and 5 show examples of feature matching and affine transformations by LoFTR.

2) Preprocessing to Reduce Domain Gap: Table III shows the inference result obtained after preprocessing SWIR images by inverting their intensity and denoising. With LoFTR, the success rate (considering success as within 10 pixels) after applying preprocessing was improved by 21% compared to that without applying preprocessing. With SiLK, the success remained 0% after applying preprocessing. These results demonstrate that applying the preprocessing steps to align the features of both IR images can improve the success rate. Therefore, we determined that SiLK is unsuitable for our task. Figures 6 and 7 show an example of the inference result obtained using LoFTR after preprocessing the SWIR image by inverting its intensity and denoising.

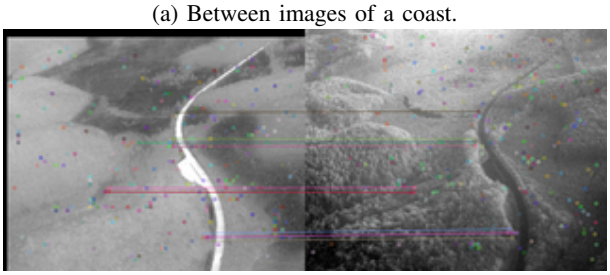
3) Data Augmentation for Model Robustness: Table IV shows the results obtained when using MegaDepth with data augmentation for training LoFTR. The success rate (consid-

TABLE II: Comparison of SiLK and LoFTR. #matches shows average number of matches per frame.

Model	Affine est. Acc			#matches
	@10px	@30px	@50px	
LoFTR (MegaDepth)	19.60	33.99	44.18	158.73
SiLK (top-10k)	0.00	0.00	0.00	0.94

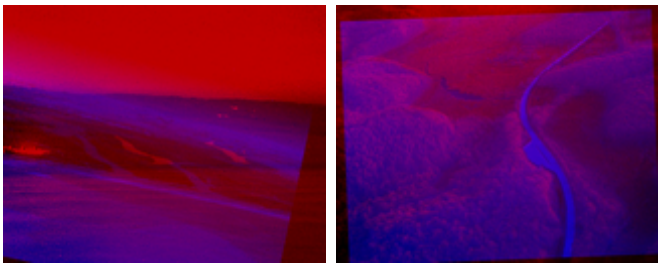


(a) Between images of a coast.



(b) Between images of a river.

Fig. 4: Examples of feature matching using LoFTR.

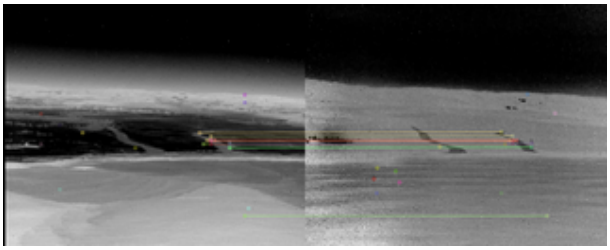


(a) Image of a coast. (b) Image of a river.

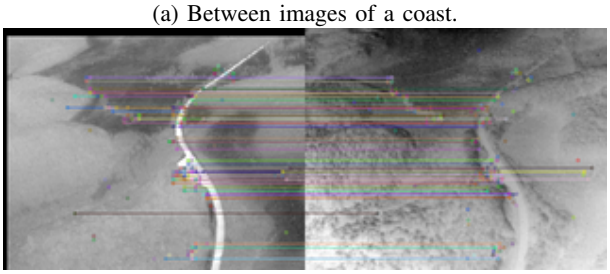
Fig. 5: Examples of image registration using LoFTR.

TABLE III: Validation of preprocessing.

Model	Eval Dataset	Affine est. Acc		
		@10px	@30px	@50px
LoFTR (MegaDepth)	raw	19.60	33.99	44.18
LoFTR (MegaDepth)	invert denoise	41.08	51.60	55.59
SiLK (top-10k)	invert denoise	0.00	0.00	0.00

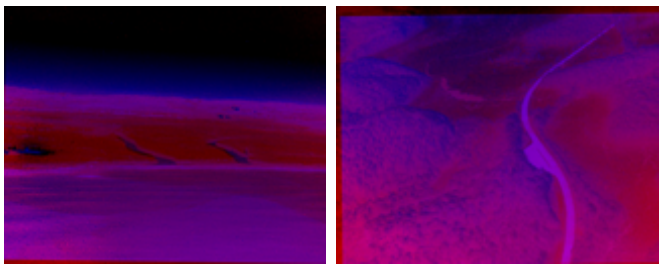


(a) Between images of a coast.



(b) Between images of a river.

Fig. 6: Examples of image registration using LoFTR for feature matching after preprocessing the SWIR by inverting its intensity and denoising.



(a) Image of a coast. (b) Image of a river.

Fig. 7: Examples of image registration using LoFTR for feature matching after preprocessing the SWIR by inverting its intensity and denoising.

ering success as within 10 pixels) of LoFTR trained with augmented MegaDepth was improved by 26% compared to that of LoFTR trained with simply MegaDepth. This confirms that expanding the training data to have diverse hues improves the robustness of feature matching. Figures 8 and 9 show examples of the inference result obtained using LoFTR trained on augmented MegaDepth.

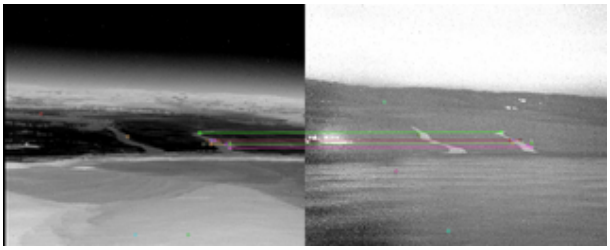
Table V shows the results obtained by performing matching and registration with LoFTR trained on the augmented MegaDepth dataset after inverting the intensity and denoising SWIR inference images. As in Table IV, the model using data augmentation showed a higher success rate of registration. Figures 10 and 11 show examples of the inference results obtained by performing matching and registration with LoFTR trained on the augmented MegaDepth dataset after inverting the intensity and denoising SWIR inference images.

V. Conclusion

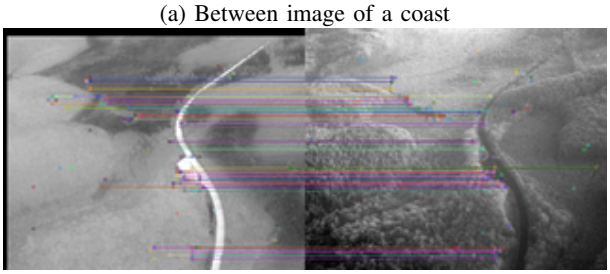
In this study, we attempted to perform registration between SWIR and MWIR images collected in a previous study [15]. Images acquired using SWIR and MWIR sensors have different characteristics, resulting in differences in the accuracy

TABLE IV: Result obtained using LoFTR with and without augmentation

Model	Affine est. Acc		
	@10px	@30px	@50px
LoFTR (MegaDepth)	19.60	33.99	44.18
LoFTR (Augmented)	46.84	59.80	61.79

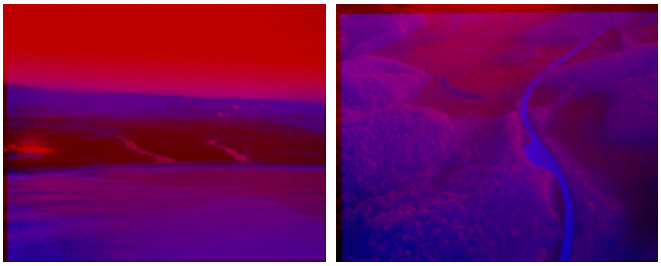


(a) Between image of a coast



(b) Between image of a river

Fig. 8: Examples of feature matching using LoFTR trained on augmented MegaDepth.



(a) Image of a coast (b) Image of a river

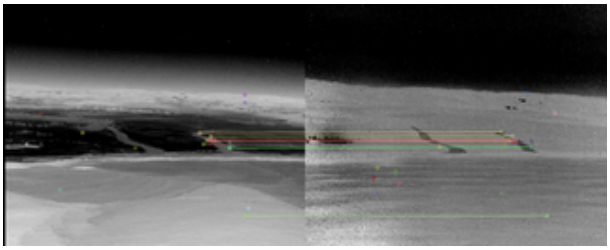
Fig. 9: Examples of image registration using LoFTR trained on augmented MegaDepth for feature matching.

TABLE V: Result obtained using loftr with and without augmentation after inverting intensity and denoising

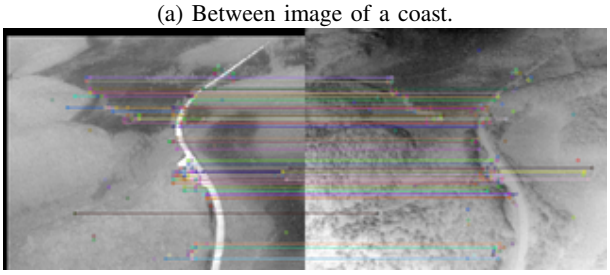
Model	Affine est. Acc		
	@10px	@30px	@50px
LoFTR (MegaDepth)	41.08	51.60	55.59
LoFTR (Augmented)	57.47	72.64	74.86

of the feature matching applied for registration. To mitigate this problem, we investigated various image processing and deep learning methods for improving the feature matching between SWIR and MWIR images.

Initially, we compared the efficacy of LoFTR and SiLK to identify whether a detector-based or a detector-free feature matching method would be more suitable for our task. We found that LoFTR achieved a registration success rate of 44% on unprocessed SWIR and MWIR images, whereas SiLK had a 0% success rate. These results corroborate the conventional

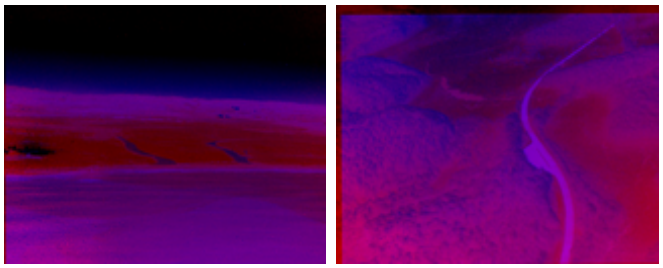


(a) Between image of a coast.



(b) Between image of a river.

Fig. 10: Examples of feature matching using LoFTR trained on augmented MegaDepth after preprocessing the SWIR image by inverting its intensity and denoising.



(a) Image of a coast. (b) Image of a river.

Fig. 11: Examples of image registration using LoFTR trained on augmented MegaDepth for feature matching after preprocessing the SWIR image by inverting its intensity and denoising.

notion that detector-based methods like SiLK are susceptible to domain gaps, such as changes in lighting conditions.

SWIR and MWIR images exhibit the phenomenon of reversed black and white attributes according to Kirchhoff's law. In addition, SWIR images are prone to noise. Consequently, by inverting the intensity of and denoising SWIR images before performing feature matching, we were able to increase the registration success rate to 55%.

Further, we applied data augmentation to diversify the hues in our training data. As a result, the success rate improved further to 74%. This result indicates that data augmentation improves the matching accuracy, thus demonstrating the efficacy of this approach.

REFERENCES

[1] K. Chrzanowski, "Review of night vision technology," *Opto-Electronics Review*, vol. 21, no. 2, pp. 153–181, 2013.

[2] Rogalski Antoni, "Next decade in infrared detectors," in *Proc. SPIE-Int. Soc. Opt. Eng*, 2017.

[3] Sandhya Kumari, Sanagapallea Koteswara Rao, and I. Prabha, "Contrast enhanced low-light visible and infrared image fusion," *Defence Science Journal*, vol. 66, pp. 266–271, 04 2016.

[4] Hui Li and Xiao-Jun Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, pp. 2614–2623, 12 2018.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, nov 2004.

[6] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[7] Pablo Fernndez Alcantarilla, Adrien Bartoli, and Andrew Davison, "KAZE features," in *Proc. ECCV*, 10 2012.

[8] Pablo Fernndez Alcantarilla, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *British Machine Vision Conference*, 09 2013.

[9] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[10] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.

[11] Pierre Gleize, Weiyao Wang, and Matt Feiszli, "SiLK–simple learned keypoints," arXiv preprint arXiv:2304.06194, 2023.

[12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[13] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy, "Domain generalization: A survey," *Proc. IEEE Trans. Pattern Anal. Mach. Intell*, vol. 45, no. 4, pp. 4396–4415, 2023.

[14] M. Adachi, H. Komatsuzaki, M. Wada, and R. Miyamoto, "Accuracy improvement of semantic segmentation trained with data generated from a 3d model by histogram matching using suitable references," in *Proc. IEEE SMC*, 2018, pp. 1180–1185.

[15] M. Kibe, Y. Kishi, S. Kobayashi, and J. Kudo, "Image characteristics of SWIR and MWIR type-ii superlattice detector arrays for night vison applications," *The journal of the Institute of Image Information and Television Engineers*, march 2023.

[16] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[17] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, 2007.

[18] A. Toet, "The TNO multiband image data collection," *Data in Brief*, 2017.

[19] F. B. Campo, F. L. Ruiz, and A. D. Sappa, "Multimodal stereo vision system: 3d data extraction and algorithm evaluation," *IEEE Journal of Selected Topics in Signal Processing*, 2012.

[20] F. Barrera, F. Lumbreras, and A. D. Sappa, "Multispectral piecewise planar stereo using manhattan-world assumption," *Pattern Recognition Letters*, 2013.

[21] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, 2012.

[22] Matthew Brown and Sabine Ssstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 177–184.

[23] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in Computer Vision*, 1987.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[25] K. M. Yi, E. Trulls, V. Lepetit, and P. FuaGiang, "LIFT: Learned invariant feature transform," in *Proc. ECCV*, 2016.

[26] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Toward geometric deep SLAM," arXiv:1707.07410, 2017.

[27] K. T. Giang, S. Song, and S. Jo, "TopicFM: Robust and interpretable topic-assisted feature matching," arXiv:2207.00328, 2022.

[28] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "MatchFormer: Interleaving attention in transformers for feature matching," in *Proc. Asian Conf. on Comput. Vis.*, 2022.

[29] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[30] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft COCO: Common objects in context," 2015.

[31] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.