Automatic Annotation Method for Day-Night Aerial Infrared Image Dataset Creation and Its Application to Semantic Segmentation

Michiya Kibe, Takeru Inoue, Junya Morioka, and Ryusuke Miyamoto,

Abstract—In situations where visible lighting is inadequate for sensing, infrared sensors are commonly employed. However, they often yield blurry images lacking clear textures and terrain/object boundaries. Unfortunately, human visibility diminishes even with infrared sensors providing more visual information than visible sensors, especially at night aerial imagery. To enhance the visibility of aerial infrared images, we propose adopting semantic segmentation, which assigns pixel-wise class labels to various input images, thereby clarifying substantial boundaries. However, training an accurate semantic segmentation model necessitates extensive pixel-wise annotations corresponding to input images, which are lacking in aerial infrared images with ground truth datasets. To address this challenge, we introduce a novel method that automatically generates pixel-wise class labels using solely infrared images and metadata such as GPS coordinates. Our method comprises two pivotal functions: coarse alignment with metadata in geographic information system (GIS) space and fine alignment based on multimodal image registration between aerial images. Aerial image datasets spanning three domains—day, twilight, and night—were created using short-wave infrared (SWIR) and mid-wave infrared (MWIR) images captured by optical sensors mounted on helicopters. Experimental results demonstrate that training on GIS data as label images enables high-precision semantic segmentation across both daytime and nighttime conditions.

Index Terms—Semantic segmentation, Automatic annotation, Multimodal image registration, Aerial infrared image

1 INTRODUCTION

HELICOPTERS and UAVs are indispensable in numerous tasks such as security, monitoring, and search and rescue operations, where visibility is paramount. In situations where visible wavelengths are inadequate for sensing, infrared sensors are commonly utilized. However, they often produce blurry images lacking clear textures and boundaries between different terrains and objects. Unfortunately, human visibility worsens at night despite infrared sensors providing more visual information than visible sensors. We propose adopting semantic segmentation to address this issue and enhance the visibility of aerial infrared images. This technique assigns pixel-wise class labels to various input images, clarifying substantial boundaries. In fields like autonomous driving, remarkable performance has been achieved through various semantic segmentation models for environmental perception [1], [2]. If these advanced models prove effective for aerial infrared images, they could substantially enhance situational awareness capabilities during aerial missions throughout the day and night.

Training an accurate model for semantic segmentation necessitates extensive pixel-wise annotations corresponding to input images. Though manual labeling is commonly used to obtain labeled images, it presents challenges due to its expense and difficulty in labeling night images with unclear boundaries [3]. Several methods for obtaining labeled images have been deployed to overcome these issue. In the SODA dataset [4], the pix2pixHD [5] model is utilized to translate thermal images from labeled RGB images, thereby expanding the dataset size. Similarly, the Freiburg Thermal (F–T) dataset [6], which contains only 32 labeled images for validation, employs the "HeatNet" model. Heatnet generates teacher images from RGB day images and employs

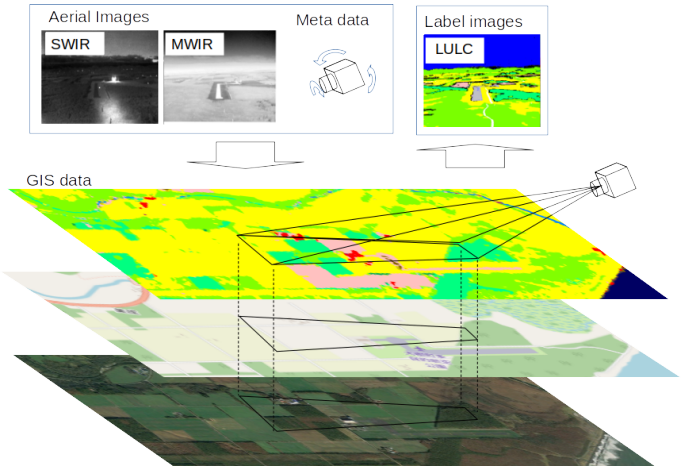

Fig. 1. Our annotation method utilizes geographic information system (GIS) data, enabling the automatic generation of pixel-wise class labels solely from infrared images and metadata such as GPS coordinates. This approach facilitates training on images across three distinct domains: day, twilight, and night, enabling the implementation of high-precision semantic segmentation across varying lighting conditions from day to night.

- *M. Kibe, T. Inoue and J. Morioka are with Department of Computer Science, Graduate School of Science and Technology Meiji University.*
- *R. Miyamoto is with Department of Computer Science, School of Science and Technology Meiji University.*
  1-1-1 Higashimita, Tama-ku, Kawasaki, Japan.
  E-mail: {kibe, takeru, mjun, miya}@cs.meiji.ac.jp

domain adaptation for day-night transitions, improving segmentation accuracy at night. Despite the effectiveness of image translation and domain adaptation techniques, they do not eliminate the need for manual labeling. In aerial images with distant scenes, atmospheric propagation leads to substantial contrast reduction, posing challenges for manual labeling.

In contrast to previous research efforts, our proposed annotation approach leverages open GIS databases. In today's context, substantial information, including maps and satellite imagery, is available as GIS data. A study [7] successfully extracted road and building areas from OpenStreetMap (OSM) and conducted semantic segmentation using this information as labeled images aligned geometrically with visible images. This approach's effectiveness was validated through a comparison with manual labeling. If accurate georeferencing of infrared images is achievable, similar semantic segmentation of aerial infrared images can be conducted by generating labeled images from GIS databases.

Our proposed method consists of two primary functions: coarse alignment with metadata such as GPS to extract high-resolution visible images from GIS data and fine alignment based on multimodal image registration between aerial and GIS images. Traditional learning-based multimodal image registration demands many aligned multimodal image pairs for training to achieve high precision [8]. The collection of these training samples is challenging due to alignment errors caused by physical vibrations and sensor noise. Notably, we discovered that a straightforward preprocessing method based on physical mechanisms effectively mitigates multimodal gaps between images. Our findings demonstrate that multimodal image registration is achievable without additional training, utilizing a model trained solely on visible images. This approach enables the creation of task-specific labeled images from extensive open GIS information.

For the construction of our aerial image dataset, we leverage previous research data [9], which comprises images captured from a helicopter using two infrared sensors: short-wave infrared (SWIR) and mid-wave infrared (MWIR). These images exhibit diverse characteristics based on time domains, including day, twilight, and night. Our study demonstrates the feasibility of training models capable of achieving high-precision semantic segmentation throughout the day and night using labeled images from GIS databases.

The main contributions of this study are as follows:

- Proposal of an automatic annotation method utilizing open GIS databases: Our method eliminates the need for costly manual labeling, enabling the acquisition of labeled images even for challenging night images.
- Application of a simple preprocessing method to mitigate gaps between images and facilitate multimodal image registration without requiring additional training: This approach enables high-precision georeferencing of day-night aerial infrared images.
- Creation of aerial image datasets encompassing three domains—day, twilight, and night—utilizing SWIR and MWIR images: These datasets serve as valuable research resources.

Demonstration through experiments using the created datasets that training a single model on all domains enables high-precision semantic segmentation throughout both day and night scenarios. These contributions advance the field by providing an efficient and practical approach to annotating aerial infrared images, enabling precise georeferencing, and achieving consistent semantic segmentation performance across varying lighting conditions.

2 Related Work

2.1 Infrared sensor for night vision

Infrared wavelengths are classified into various bands, including near-infrared (NIR, 750 nm–1 µm), SWIR (1–2.5 µm), MWIR (3–5 µm), and long-wave infrared (LWIR, 8–12 µm). NIR and SWIR are employed in low-light vision sensors that utilize moonlight or nightglow as light sources. In contrast, MWIR and LWIR are utilized in thermal imaging sensors that detect the radiation emitted by objects themselves.

Image intensifiers (II) utilizing NIR are widely employed as low light level sensors [13]. Additionally, sensors utilizing avalanche effects, like EBAPS [14] and SPAD [15], have advancements in performance. Recently, susceptible SWIR sensors with excellent dehazing capabilities, particularly for phenomena like smoke, are gaining prominence [13], [16].

Thermal imaging sensors can be categorized into thermal detectors (bolometers) and quantum detectors (photovoltaic and photoconductive detectors). Thermal detectors are uncooled, making them cost-effective, while quantum detectors, cooled to around 80K–140K, offer higher sensitivity than thermal detectors. Thermal detectors typically utilize LWIR, while quantum detectors can also utilize MWIR through semiconductor bandgap control [17].

In numerous image processing studies, images from NIR and LWIR are commonly employed. Due to their atmospheric propagation properties, SWIR and MWIR sensors are particularly advantageous for their superior contrast characteristics in capturing distant scenes. This makes them especially relevant in aircraft applications, where longer ranges are essential [18], [19]. Despite their advantages, there is a call for further advancement in the study of image processing tailored explicitly for these wavelength bands–SWIR and MWIR.

2.2 Day-night infrared image datasets and semantic segmentation

Several publicly available datasets contain infrared images [20], some of which include day-night images. Representative day-night image datasets relevant to our research are summarized in Table 1. The MSOD [10] dataset focuses on object detection, such as pedestrians, and comprises images across a wide range of wavelengths. The SODA [4] and Freiburg Thermal [6] datasets employ dataset expansion techniques through image translation and domain adaptation methods to address the challenge of annotating complex night images, respectively. These datasets are instrumental in achieving excellent semantic segmentation accuracy using their original models. The MVSeg [3] dataset

TABLE 1
Relevant publicly available day-night infrared image datasets.

Dataset attributes		MSOD [10]	SODA [4]	F-T [6]	HIT-UAV [11]	MVSeg [3]	MONET [12]	Ours
Type	Terrestrial	✓	✓	✓		✓		
	Aerial				✓		✓	✓
Sensor Modality		RGB, SWIR, MWIR, LWIR	LWIR	RGB, LWIR	LWIR	RGB, LWIR	LWIR	SWIR, MWIR
Task	Detection	✓			✓		✓	
	Segmentation		✓	✓		✓		✓
Classes		-	20	12	-	26	-	11
# annotated images: manual(not manual)		30K	2,168 (5,000)	32 (20K)	2.9K	3,545	53K	(2,940)
Year		2017	2019	2020	2022	2023	2023	-

gathers existing public datasets for object detection tasks, such as RGBT234 [21] and KAIST [22], and undertakes manual annotation to create a new RGB-T benchmark dataset for semantic segmentation.

While many infrared image datasets have centered on traffic scenes, there has been a surge in aerial image datasets in recent years. The HIT–UAV [11] and MONET [12] datasets consist of day-night images captured using LWIR cameras mounted on UAVs in urban and rural areas. Although these datasets are primarily designed for detection tasks such as pedestrians and vehicles, to our knowledge, there is currently no aerial infrared image dataset especially tailored for semantic segmentation.

In a previous study [9] conducted by the author, a dataset was created by capturing images of various forward-looking scenes using two infrared sensors, SWIR and MWIR, mounted on a helicopter . The infrared sensors, as shown in Fig. 2, are separately mounted on gimbals, controlled to align with the same direction as the HMD (Helmet-Mounted Display) inside the aircraft. Although the dataset includes various scenes such as vegetation, water, roads, runways, and the sky, creating pixel-wise labeled images is necessary for use in semantic segmentation. In this study, we utilize this image dataset to demonstrate automatic annotation.

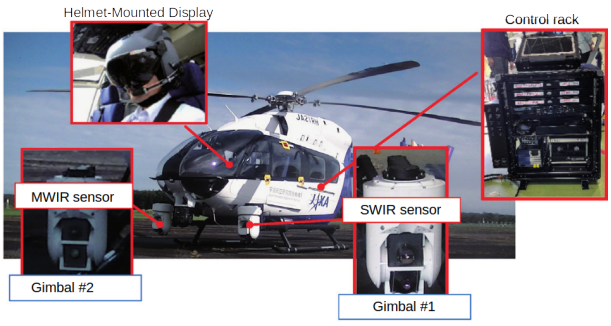

Fig. 2. The two infrared sensors, SWIR and MWIR, are mounted on separate gimbals attached to the left and right steps of the helicopter.

2.3 Georeferencing and multimodal image registration

Georeferencing is the process of associating data, such as maps or aerial images, with specific geographic locations on the Earth's surface. It can be utilized to overlay raster images of various GIS sources and obtain information for specific geographic location points. Georeferencing of aerial images to geographic locations can be achieved through coordinate transformations using metadata like GPS and orientation sensors. Sensor noise and platform vibration cause alignment error. Pixel-level fine alignment is achieved through image registration between aerial image and GIS image [23].

In recent years, deep learning-based feature detection and matching methods have been extensively developed for image registration, demonstrating better performance than hand-craft ones. Notable examples include SuperPoint [24], which detects local features and outputs their descriptors using CNN, and SuperGlue [25], which directly outputs correspondences between sets of features using a graph neural network. LoFTR [26] is another method that extracts features using CNN and Transformer, performing end-to-end tasks from detection to matching. While these methods enable precise alignment between single modal images, ones between multimodal images remains challenging. For instance, Multipoint [8] which uses the architecture of SuperPoint addresses the multimodal image registration by learning from aligned visible and infrared image pairs.

In cases where there are not enough sufficient image pairs for training, preprocessing to mitigate the modal gap between images can be effective. In [27], image-to-image translation using GAN model to convert infrared images into visible images is combined with Multipoint. Additionally, [28] combines preprocessing that extracts modality-invariant geometric structural information using Log-Gabor filters with LoFTR. Moreover, our previous work [29] conducted experiments using augmentation and preprocessing tailored to each image's characteristics with LoFTR, utilizing the image dataset of [9]. This demonstrated the high-precision multimodal registration between SWIR and MWIR images.

3 Proposed Automatic Annotation Method

In this section, we introduce our proposed annotation method that utilizes open GIS databases, automatically generating labeled images using only aerial infrared images and metadata as input information.

3.1 Overview

The flow for creating labeled images and datasets is illustrated in Fig. 3. Our method consists of two major functions.

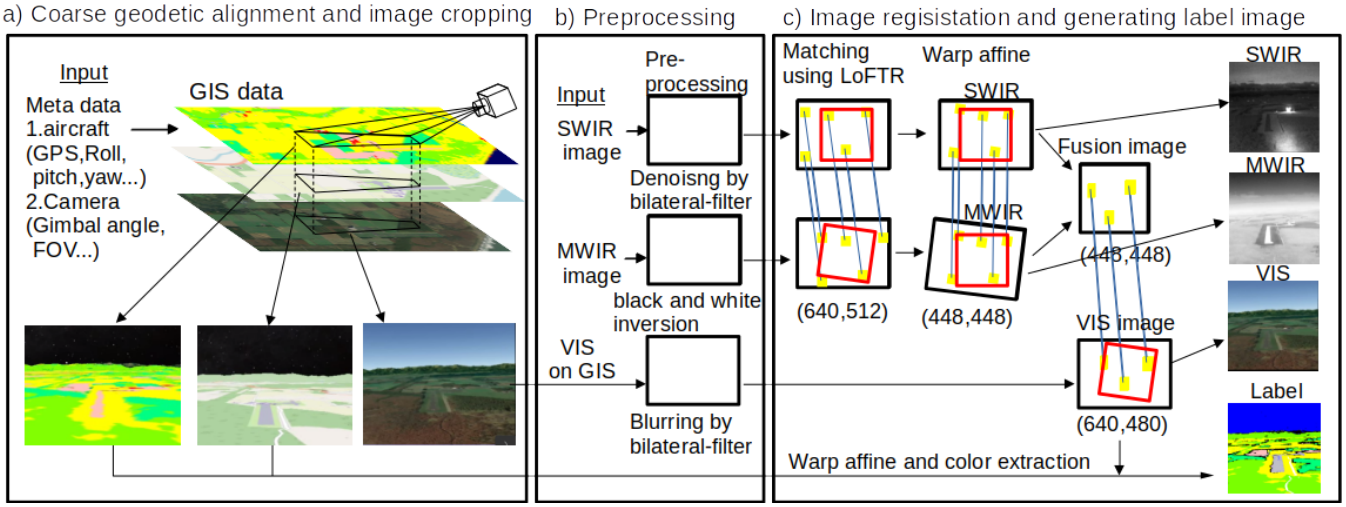

Fig. 3. Flow for creating label images and datasets, which involves a two-step approach: coarse alignment, depicted in process a), fine alignment consists of processes (b) and (c).

Firstly, coarse alignment is performed using metadata such as GPS to crop projected images on the GIS space. Next, fine alignment is conducted through multimodal image registration, creating a dataset containing labeled images using the obtained coordinate transformation matrix. Additionally, preprocessing tailored to the characteristics of infrared images is applied in this method to perform image registration without additional training.

3.2 Coarse geodetic alignment and image cropping

The purpose of coarse geodetic alignment is to crop GIS images with an equivalent field of view as aerial infrared images, serving as a preliminary step for the fine alignment described next subsection. The geometric transformation from the world coordinates to the camera coordinates is geometrically determined by the following equations, utilizing eight types of metadata: 3 degrees of freedom of aircraft attitude, 2 degrees of freedom of camera pose, and 3 degrees of freedom of aircraft position, as described below.

$$T = \begin{pmatrix} 1 & 0 & 0 & \delta T_x \\ 0 & 1 & 0 & \delta T_x \\ 0 & 0 & 1 & \delta T_x \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{1}$$

$$R = \begin{pmatrix} \cos\phi & 0 & -\sin\phi & 0 \\ 0 & 1 & 0 & 0 \\ \sin\phi & 0 & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\beta & \sin\beta & 0 \\ 0 & -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2}$$

$$G = \begin{pmatrix} \cos\omega & 0 & -\sin\omega & 0 \\ 0 & 1 & 0 & 0 \\ \sin\omega & 0 & \cos\omega & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\tau & -\sin\tau & 0 & 0 \\ \sin\tau & \cos\tau & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

and

$$\overrightarrow{X}_{camera} = G \cdot R \cdot T \cdot \overrightarrow{X}_{world}, \tag{4}$$

where T is the rotation matrix related to the camera's elevation angle (ω) and azimuth angle (τ), G is the rotation matrix related to the aircraft's pitch angle (ϕ), roll angle (β), and heading angle (α), and T is the translation matrix related to the aircraft's position (latitude, longitude, altitude). By operating these matrices in GIS space, it is possible to project onto the same Field of View (FOV) as aerial images. Due to temporal and spatial noise in the metadata used for these calculations, only coarse alignment is achievable at this stage.

For semantic segmentation, accurate pixel-wise alignment is required. Although infrared multispectral images captured from satellite are available as GIS data today, the present ground resolution is typically in the order of several tens of meters, lower than that of aerial infrared images. Applicable images on GIS are limited to high-resolution visible images. The Google Earth platform is an online virtual geographic information system software that provides high-resolution visible images with a resolution exceeding one meter for most scenes worldwide. In this study, we utilize this platform to crop the visible images needed for image registration and to obtain the GIS images required for label creation.

3.3 Preprocessing for modal gap adaptation

Here, we introduce the preprocessing method for our proposed modal gap adaptation. Unlike single-modal images, feature matching between multimodal images with differences of spectra remain challenges. When the spectra differ, differences in edge characteristics, noise characteristics, and intensity polarity occur for each modality. Assuming the Earth's surface is Lambertian and ignoring the influence of atmospheric propagation, the radiance is expressed by the following equation:

$$L_0(\lambda) = \epsilon(\lambda)B(T,\lambda) + r(\lambda)\frac{F_0(\lambda)}{\pi}, \tag{5}$$

where λ, T, ϵ and r represent the wavelength, surface temperature, spectral emissivity, and spectral reflectance of the Earth's surface, respectively. B and F_0 denote the blackbody spectral radiance of the surface temperature and the downward irradiance illuminating from external light, respectively. The first term represents the thermal radiation component of the object itself, and the second term represents the reflected component of the incident light from the external source. The differences in image characteristics arise due to the modalities of detection wavelengths and the domains of day-night.

During day, there is radiance from sunlight across all wavelength bands, and each image contains some degree of reflected components, making it easy to extract common features. At night, the differences between modalities become more prominent. SWIR exhibits reflective radiance from moonlight and night glow, while MWIR is dominated by the thermal radiation emitted by objects themselves, making feature matching challenging.

We notice that by applying preprocessing tailored to each wavelength band based on physical mechanisms such as reflection, radiation, and imaging conditions, it is possible to mitigate modal gaps between images. Table 2 summarizes the preprocessing techniques adopted to align the characteristics of each image. For SWIR images, a bilateral filter with high edge-preserving capability [30] is employed to reduce noise in low-light conditions. For MWIR images, unlike other modalities at night, which have no reflective components, we adopt a process of black and white inversion. The relationship between reflectance (r) and emissivity (ϵ) is expressed by the following equation based on the laws of energy conservation and Kirchhoff's law:

$$\epsilon(\lambda) = 1 - r(\lambda), \tag{6}$$

According to this equation, the radiance of regions with high reflectance is high during day, while it is low during night, and vice versa. Black and white inversion makes the radiance characteristics of the thermal image similar to those of reflective images. Additionally, in the VIS image on GIS, taken at a different time, detailed information such as texture has differences from other modalities. To achieve a blurring effect for detailed information, a bilateral filter is employed. Fig.4 illustrates images before and after processing for each modality. Applying preprocessing reveals that the edge, noise, and intensity polarity characteristics of each image are becoming more similar. We demonstrate the enhanced precision of keypoint matching at night images through this preprocessing.

TABLE 2
Modality-specific image characteristics and each pre-processing selected for mitigating modal gaps.

modality time	SWIR night	MWIR night	VIS(on GIS) day
sharpness noise image property	good very noisy reflective	poor low radiative	very good low reflective
preprocess	bilateral filter	black and white inversion	bilateral filter

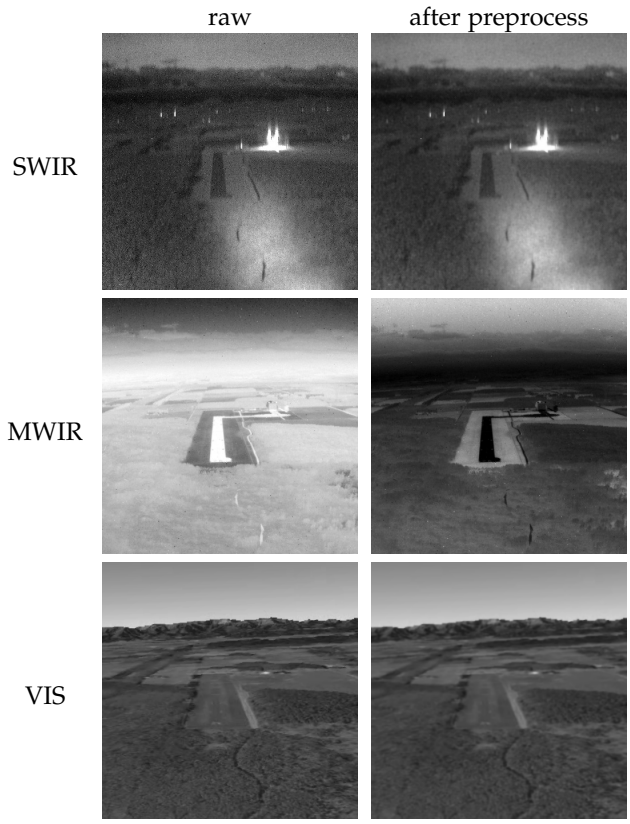

Fig. 4. Visible and Infrared Images Before and After Preprocessing

3.4 Image registration and generating label images

We apply preprocessing to only VIS images in day and all modal images in twilight and night, performing keypoint matching using LoFTR [26]. LoFTR is a detector-free approach that utilizes CNN and attention mechanisms to produce robust matching results. In contrast to detector-based methods employing handcrafted feature extractors like SIFT [31], SURF [32], or deep learning-based models like MultiPoint, LoFTR leverages long-range context to obtain high-quality correspondences, even in regions with low texture.

Here, we use the outdoor model trained on MegaDepth [33] with dual softmax matching in LoFTR, without additional training. The matching points obtained by LoFTR are used to estimate the coordinate transformation parameters through the RANSAC algorithm [34]. Considering the relatively minor influence of parallax in this scenario, an affine transformation is used with fewer parameters to minimize computational errors. Although affine transformation parameters can be estimated with three or more corresponding points, we consider image pairs with 20 or more inliers as correctly estimated due to the decrease in accuracy with fewer points.

Alignment with visible images involves utilizing a fusion image of SWIR and MWIR to leverage features from both modalities. Initially, multimodal image registration between SWIR and MWIR images is performed, followed by blending the two images in equal proportions using $/alpha$-blending. Subsequently, multimodal image registration between VIS images and the fusion images is conducted.

All GIS images required for the label image are obtained by applying the affine transformation matrix obtained in this process and cropping them to have the same FOV as the infrared images. Finally, label images are generated by extracting class regions using color information from all GIS images.

4 Experimental Result

In this section, we present the creation of datasets using the proposed method, experimental results on semantic segmentation, Quantitative evaluation by various learned based models, and ablation study on multimodal image registration.

4.1 Creation of Datasets Using the Proposed Method

This study utilized Google Earth, OpenStreetMap, and the land-use and land-cover map from the Japan Aerospace Exploration Agency (JAXA) [35] to perform semantic segmentation for the purpose of piloting assistance. Additionally, we considered the following factors when creating the aerial dataset:

- Humans take time for dark adaptation, accidents due to decreased visibility are expected to increase in twilight [36]. The accuracy of classification during this period is also crucial.
- Existing studies often focus on NIR and LWIR images. SWIR and MWIR are also widely used [18], [19]. Research using these wavelength bands, similar to NIR and LWIR, needs to be further advanced.

The dataset utilized images and metadata from the author's previous study [9], creating three domains: day, twilight, and night. To enable learning from landscapes as similar as possible in each domain, we selected scenes primarily involving flight over flat areas and approach to runways. Fig.5 demonstrates an example of applying the proposed method to feature point matching in a night image. Inlier correspondence points are shown in blue lines, and outlier outliner points are indicated by red lines. In this scene, more than 20 inlier correspondence points are calculated, indicating successful matching. the detail of matching performance is described in Subsection D, E.

We applied the proposed method to scenes in all domains and created a dataset. The dataset consists of 726 images for the day domain, 1713 images for the twilight domain, and 501 images for the night domain. Example images of the dataset is illustrated in Fig. 6. Table 3 shows the categories, colors for each class and sample counts (80 frames) in the test images for each domain. The scene locations primarily consist of rural areas, resulting in a dataset with a somewhat lower number of built-up. There is not a significant imbalance in the sample counts of classes in each domain, making it a suitable configuration for domain comparison experiments.

4.2 Experimental Results of Semantic Segmentation

The objective of this experiment is to verify whether semantic segmentation can be achieved consistently throughout the day and night using label images created from

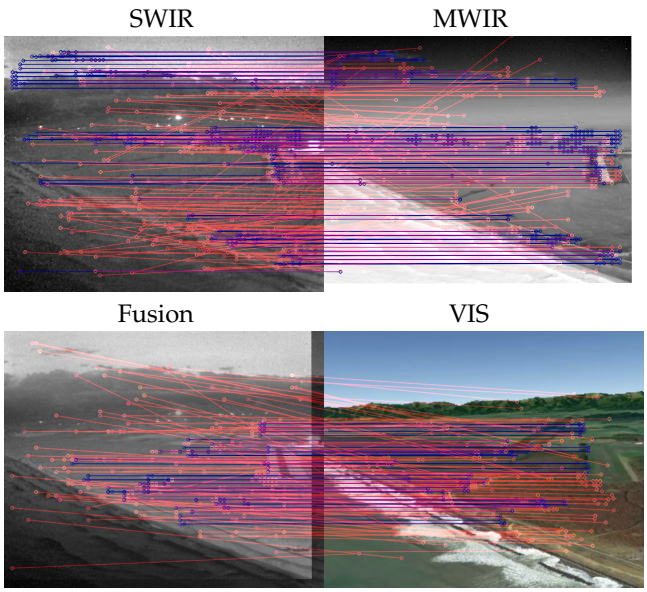

Fig. 5. Example of keypoint matching using LOFTR: The blue lines indicate inliers, while the red lines indicate outliers.

TABLE 3
Information of each classes in the created dataset. "num of pix" shows the number of samples in each domain of the test images (80 frames).

No.	Color	Category	Num of Pix		
			Day	Twilight	Night
0		None	1,136,374	1,353,939	821,131
1		Sky	2,402,448	5,204,791	2,294,247
2		Built-up	5,464	67,948	12,412
3		Paddy field /Grassland	6,180,552	4,844,902	7,486,231
4		Cropland	260,679	310,539	86,270
5		DBF	3,708,962	2,537,218	3,228,186
6		DNF	405,531	450,037	117,247
7		ENF	71,388	82,652	104,046
8		Bare	65,963	60,595	76,686
9		Water	1,529,865	980,600	1,620,345
10		Road	110,426	84,363	93,544
11		Runway	178,668	78,736	115,975

GIS data. For training and validation datasets, 320 and 80 images were randomly selected from the day, twilight, and night domains, respectively. As a baseline, we used the PSPNet [37] with ResNet50 [38] as the backbone. The input consisted of pseudo-RGB images with three channels, including 2 channels for SWIR and MWIR, and one channel filled with zeros. The models were trained for 200 epochs, and evaluation was performed using test images. The initial weights were derived from a pre-trained model on the Cityscapes visible image dataset.

First, we evaluated the models trained on specific domains, followed by an evaluation using models trained on all domains. The results are presented in Table 4, and Fig. 7 shows the segmentation images inferred by each model. While high performance was obtained when the domain for training and testing was the same, significant misclassifications occurred when the domains were different. On the other hand, the model trained on all domains achieved consistently high scores across all domains. This result confirms that training across multiple domains enables high-precision segmentation throughout the day and night using the same

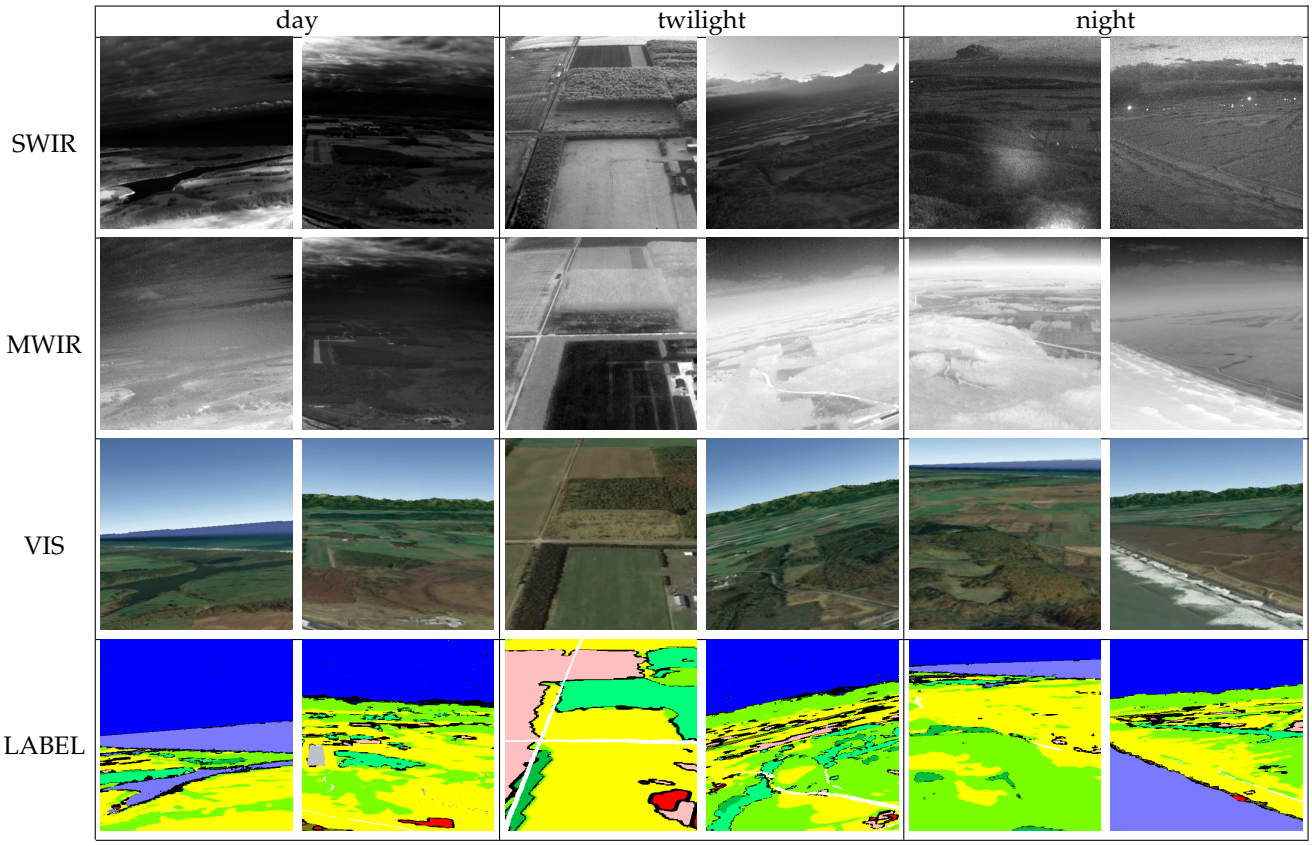

Fig. 6. Example images from the created dataset showcasing scenes from three domains: day, twilight, and night. The dataset encompasses a variety of rural landscapes, including sea, river, forests, roads, and runways.

model.

TABLE 4
Comparison of scores from models trained on each domain and all domains.

Model/Backbone	train on	test on	mIoU	aAcc	mAcc
PSPNet/ Resnet50	day	day	57.4	90.4	67.7
		twilight	8.8	43.8	20.2
		night	9.9	49.6	16.2
		all	24.6	61.3	35.9
	twilight	day	9.9	37.3	15.0
		twilight	58.0	90.0	69.6
		night	34.9	82.2	43.6
		all	34.5	69.9	45.6
	night	day	9.8	37.8	17.7
		twilight	23.7	67.9	35.4
		night	63.8	94.2	74.8
		all	29.2	66.8	40.7
	all	day	56.4	88.8	68.4
		twilight	58.4	90.1	69.3
		night	64.7	93.9	74.2
		all	**61.9**	**90.9**	**73.1**

4.3 Quantitative evaluation

Here, we conducted quantitative evaluation using the created image dataset. Validation experiments using various visible image datasets have already been conducted, demonstrating the effectiveness of semantic segmentation by many models, including CNNs and Transformers, achieving high classification accuracy. We performed training and evaluation using representative models, CNN-based DeeplabV3+ [39], PSPnet [37], and Transformer-based Mask2Former [40] and Segformer [41].

The experiment results are presented in Table 5. Among the CNN-based models, PSPNet (Resnext101d) achieved the highest score. In the Transformer-based models, SegFormser (InternImage) demonstrated the highest score, surpassing PSPNet (Resnext50) used in the previous experiment by an improvement of over 2.5% in mean intersection over union(mIoU). Additionally, the experiments used a pretrained model for visible images as the initial weight. The evaluation results of a model trained solely on the dataset without pretraining are presented as PSPNet (NoPretrain). A comparison revealed that using pretraining with visible image models resulted in a score over 5% higher, indicating the effectiveness of pretraining visible image models even for infrared image datasets.

4.4 Comparison test of multimodal image registration

To evaluate the performance of our multimodal image registration, we conducted comparative experiments with SIFT, Superpoint and SuperGlue (SP+SG). Since the models with raw input images showed low scores, we compared the results obtained with preprocessed images. The results are presented in Table 6. Here, accurate pairs(Acc pairs) represent the average number of inliers in affine estimation among the matching points, Accuracy (Acc) is the ratio of accurate pairs to matching pairs, and precision score (Pre) indicates the percentage of images with successful matches (inliers: 20 pairs or more) out of the total samples. For the

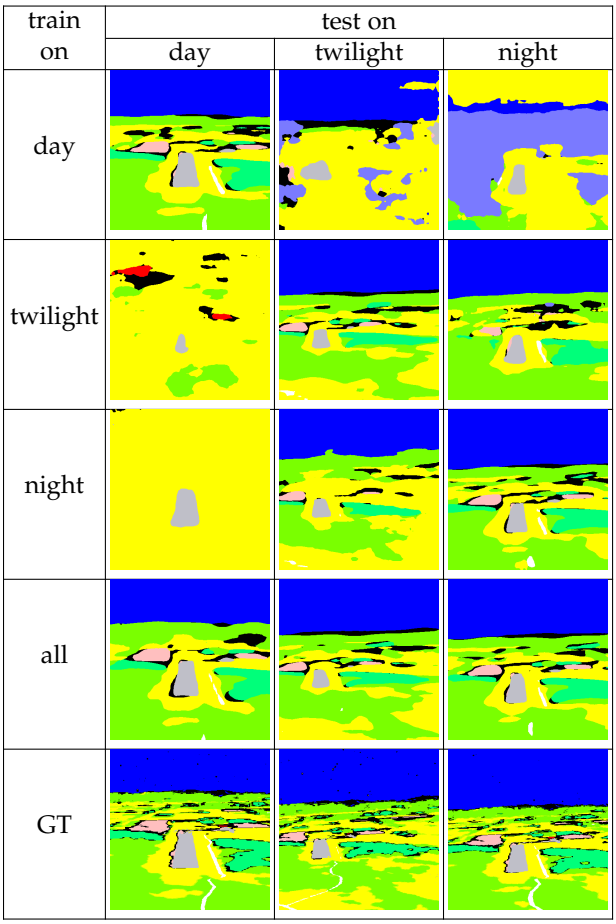

Fig. 7. Difference of the inference images based on each training domain: Training on the same domain leads to fewer misclassifications in the inference images.

TABLE 5
Comparison of scores from various models

Model	Backbone	mIoU	aAcc	mAcc
PSPNet	Resnet50	61.9	90.9	73.1
PSPNet	Resnet101	59.5	91.5	71.2
PSPNet	Resnext101 [42]	63.7	**92.3**	**77.0**
PSPNet	InternImage [43]	60.3	90.9	74.6
DeepLabV3+	Resnet50	60.4	90.6	69.4
DeepLabV3+	Resnet101	56.4	85.0	65.6
Mask2Former	Resnet50	62.8	91.1	74.9
Mask2Former	Swin-L [44]	63.4	91.0	75.3
SegFormer	MiT-B5	64.2	92.0	75.0
SegFormer	InternImage	**64.5**	**92.3**	76.0
PSPNet(NoPretrain)	Resnext101	58.6	91.2	71.0

creation of label images, the precision score is particularly crucial. In the matching of SWIR and MWIR images, SP+SG achieved the best score only during the day. In all other cases, the proposed method consistently demonstrated a high precision score. Moreover, in matching the fusion images with the reference VIS images, the proposed method obtained high precision scores in all cases. This experiment confirmed that combining modality-specific preprocessing with LoFTR results in excellent scores across all domains.

TABLE 6
Comparison of multimodal image registration performance for each model.

Modality	Time	Model	Acc Pairs	Acc	Pre
SWIR-MWIR	day	pp+SIFT	48.6	83.6%	63.4%
		pp+(SP+SG)	114.1	68.2%	**98.6%**
		ours	126.5	7.2%	92.1%
	twilight	pp+SIFT	3.3	30.9%	0.8%
		pp+(SP+SG)	55.7	54.3%	73.7%
		ours	119.7	20.5%	**90.8%**
	night	pp+SIFT	2.8	31.2%	0.8%
		pp+(SP+SG)	32.3	38.9%	61.3%
		ours	81.2	23.0%	**79.7%**
VIS-Fusion	day	pp+SIFT	1.0	16.9%	0.0%
		pp+(SP+SG)	15.0	15.7%	29.7%
		ours	48.0	15.4%	**65.6%**
	twilight	pp+SIFT	1.9	23.5%	0.5%
		pp+(SP+SG)	9.0	15.9%	10.8%
		ours	24.8	15.0%	**44.1%**
	night	pp+SIFT	1.9	28.8%	4.4%
		pp+(SP+SG)	17.1	22.5%	26.5%
		ours	56.3	20.9%	**58.3%**

4.5 Ablation study

Finally, we showed the result of an ablation study conducted to examine the effects of preprocessing on keypoint matching. This experiment focused on night domain with a significant modal gap. We systematically applied preprocessing to the images of each modality to observe the resulting effects. The scores are presented in Table 7. Without preprocessing, the highest score was achieved by the pair of visible and SWIR images, both being reflection images, with a score of 27.8%. Applying preprocessing led to a substantial improvement in scores, enhancing the success rate to 50.4%, 51.1% for both SWIR and MWIR images, respectively. Furthermore, the proposed method employs feature matching with fusion images to utilize the characteristics of both SWIR and MWIR images. This approach further improved the success rate to 58.3%.

TABLE 7
Results of the ablation study.

Modality m1-m2	Preprocess m1	Preprocess m2	Acc pairs	Acc	Pre
SWIR-MWIR			13.6	9.0%	16.0%
	✓		13.7	9.8%	19.7%
		✓	76.7	26.2%	72.4%
	✓	✓	81.2	23.0%	**79.7%**
VIS-SWIR			22.8	14.2%	27.8%
	✓		25.3	15.9%	32.0%
		✓	25.2	12.7%	39.5%
	✓	✓	40.5	18.0%	**50.4%**
VIS-MWIR			6.2	5.1%	1.2%
	✓		6.6	6.1%	0.7%
		✓	34.4	16.9%	37.3%
	✓	✓	45.1	20.3%	**51.1%**
VIS-Fusion			8.1	7.4%	5.8%
	✓		8.3	8.1%	7.7%
		✓	38.8	19.6%	39.6%
	✓	✓	56.3	20.9%	**58.3%**

5 Conclusion

In this paper, we proposed an automatic annotation method utilizing an external GIS database, with only aerial

infrared images and metadata as input. This method, employing preprocessing based on physical mechanisms, enables high-precision image registration even for blurry night images, allowing for the creation of annotated image datasets for both day and night. Using this approach, we created a day-night aerial infrared image datasets for SWIR and MWIR and conducted experiments on semantic segmentation. The results of the experiments showed excellent accuracy across day and night domains by training on images from multiple domains. Additionally, the application of state-of-the-art models effective in visible image datasets proved to be effective even for infrared images with significant modality gaps. From these results, it was demonstrated that by extracting relevant information from huge open GIS data, it is possible to train images across various domains and modalities without costly manual annotation. This dataset creation method is expected to be applicable to various tasks.

REFERENCES

[1] K. Muhammad, T. Hussain, H. Ullah, J. D. Ser, M. Rezaei, N. Kumar, M. Hijji, P. Bellavista, and V. H. C. d. Albuquerque, "Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 22694–22715, 2022.

[2] R. Miyamoto, M. Adachi, H. Ishida, T. Watanabe, K. Matsutani, H. Komatsuzaki, S. Sakata, R. Yokota, and S. Kobayashi, "Visual navigation based on semantic segmentation using only a monocular camera as an external sensor," *J. Robotics Mechatronics*, vol. 32, pp. 1137–1153, 2020.

[3] W. Ji, J. Li, C. Bian, Z. Zhou, J. Zhao, A. Yuille, and L. Cheng, "Multispectral video semantic segmentation: A benchmark dataset and baseline," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, jun 2023, pp. 1094–1104, IEEE Computer Society.

[4] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 3069–3082, 2019.

[5] 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," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2017.

[6] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8461–8468.

[7] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 6054–6068, 2017.

[8] F. Achermann, A. Kolobov, D. Dey, T. Hinzmann, J. J. Chung, R. Siegwart, and N. Lawrance, "Multipoint: Cross-spectral registration of thermal and optical aerial imagery," in *Proceedings of the 2020 Conference on Robot Learning*, 2021, pp. 1746–1760.

[9] 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.

[10] T. Karasawa, K. Watanabe, Q. Ha, A. T. d. Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017.

[11] J. Suo, T.-M. Wang, X. Zhang, H. m. Chen, W. Zhou, and W. Shi, "Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicles," *ArXiv*, vol. abs/2204.03245, 2022.

[12] L. Riz, A. Caraffa, M. Bortolon, M. L. Mekhalfi, D. Boscaini, A. Moura, J. Antunes, A. Dias, H. Silva, A. Leonidou, C. Constantinides, C. Keleshis, D. Abate, and F. Poiesi, "The monet dataset: Multimodal drone thermal dataset recorded in rural scenarios," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 2546–2554.

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

[14] L. Hirvonen and K. Suhling, "Photon counting imaging with an electron-bombarded pixel image sensor," *Sensors*, vol. 16, pp. 617, 04 2016.

[15] S. Ma, P. Mos, E. Charbon, and M. Gupta, "Burst vision using single-photon cameras," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 5364–5374.

[16] F. Rutz, R. Aidam, A. Bächle, H. Heussen, W. Bronner, R. Rehm, M. Benecke, S. Brunner, B. Göhler, P. Lutzmann, and A. Sieck, "Ingaas-based swir photodetectors for night vision and gated viewing," 10 2018, p. 2.

[17] A. Rogalski, "Next decade in infrared detectors," 10 2017, p. 100.

[18] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings, "Airborne optical and thermal remote sensing for wildfire detection and monitoring," *Sensors (Basel, Switzerland)*, vol. 16, 2016.

[19] G. G. Artan and G. S. Tombul, "The future trends of EO/IR systems for ISR platforms," in *Image Sensing Technologies: Materials, Devices, Systems, and Applications IX*, N. K. Dhar, A. K. Dutta, S. R. Babu, and K. K. Son, Eds. International Society for Optics and Photonics, 2022, vol. 12091, p. 120910B, SPIE.

[20] K. I. Danaci and E. Akagunduz, "A survey on infrared image & video sets," *Multimedia Tools and Applications*, 2022.

[21] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, pp. 106977, 2019.

[22] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1037–1045.

[23] Y. Sheikh, S. Khan, M. Shah, and R. W. Cannata, "Geodetic alignment of aerial video frames," 2003.

[24] 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.

[25] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.

[26] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[27] M. Elsaeidy, M. Emin Erkol, B. Kürşat Güntürk, and H. Fehmi Ateş, "Infrared-to-optical image translation for keypoint-based image registration," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 2022, pp. 1–4.

[28] M. Elsaeidy, M. Emin Erkol, B. Kürşat Güntürk, and H. Fehmi Ateş, "Infrared-to-optical image translation for keypoint-based image registration," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 2022, pp. 1–4.

[29] T. Inoue, M. Kibe, and R. Miyamoto, "Correspondence between swir and mwir images using augmentation and preprocessing for registration," in *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON)*. 2023, IEEE.

[30] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 839–846.

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

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

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

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

[35] S. Hirayama, T. Tadono, Y. Mizukami, M. Ohki, K. Imamura, N. Hirade, F. Ohgushi, M. Dotsu, T. Yamanokuchi, and K. N. Nasahara, "Generation of the high-resolution land-use and land-cover map in japan version 21.11," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 4339–4342.

[36] T. Iizuka, T. Kawamorita, T. Handa, and H. Ishikawa, "Refractive and visual function changes in twilight conditions," *PLOS ONE*, vol. 17, pp. e0267149, 04 2022.
[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018.
[40] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *arXiv*, 2021.
[41] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.
[42] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.
[43] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. h. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14408–14419, 2022.
[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.