IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 4, 2021, accepted November 26, 2021, date of publication November 30, 2021,
date of current version December 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131805

Accent for Visible and Infrared Registration (AVIR): Attention Block for Increasing Patch Matching Rate Through Edge Emphasis

INHO PARK[1], JONGMIN JEONG[2], AND SUNGHO KIM[1]
[1]Department of Electronic Engineering, Yeungnam University, Gyeongsan-si, Gyeongsangbuk-do 38541, South Korea
[2]Agency for Defense Development (ADD), Daejeon 34186, South Korea

Corresponding author: Sungho Kim (sunghokim@yu.ac.kr)

This research was supported by Agency for Defense Development (UD200005FD). This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

ABSTRACT In this paper, we propose an efficient attention module for visible and thermal infrared (TIR) matching deep learning networks. This method judges the right or wrong of heterogeneous sensor matching through the proposed deep learning model and increases the matching rate through the attention module using the edge-utilizing structure. This paper contributes to three aspects: 1) The first aspect is Convolutional Neural Network (CNN) structure comparisons for heterogeneous sensor registration. We consider the matching network as a classification problem when stacked heterogeneous sensor data become input of a single CNN network. Therefore, this paper shows result that is related with not only the network designed for heterogeneous sensor matching, but also various deep learning networks used for classification. 2) the second is a consideration for efficient attention module. The experiments show the module can replace lots of convolution blocks and the results achieve more better performance. The attention module uses a $1 \times k$ filter and a $k \times 1$ filter to extract horizontal and vertical edges and convolves two paths using them. 3) The third is suitable deep learning model for aerial complex visible and TIR data registration. To compare the various methods, we describe the calibration process of aerial visible and TIR data obtained directly from a drone. By using the calibrated data, this paper presents an AVIR attention block-based architecture that shows optimal matching results with minimal addition of parameters.

INDEX TERMS Visible, thermal infrared, heterogeneous sensor matching, aerial data matching, deep learning, attention module, classification.

I. INTRODUCTION

With the recent development of technology, the use and interest in Artificial Intelligence (AI) is increasing. Such AI is used for Automatic Target Recognition (ATR), Autonomous driving system, medicine, mechanics, and security. Detection and recognition using a single sensor, a common and important issue in these fields, are becoming a red ocean from a blue ocean. The limitations in the recognition and detection of a single sensor are clearly present due to the advantages and disadvantages of the sensor characteristics. In the case of the visible camera that uses the reflection band of the light spectrum, it has a high-resolution field of view and spatial resolution, but there is a limit that cannot be seen at night. In the case of Thermal Infrared (TIR), only the rough outline of an object can be checked with a low resolution, but because it uses the emitted radiance information, it has the advantage of being able to distinguish objects both day and night and obtaining thermal information of the object [1]. Our research team wanted to proceed with object detection through a heterogeneous sensor fusion network using visible and infrared (IR) data, which have different advantages. During the pre-processing of the implementation of this network, it was confirmed that the frame per second (FPS) of the two videos was different due to telecommunication and hardware limitations despite the same settings. This is a problem that occurs when two cameras are attached to the same location but acquire data with different devices. In addition, each video also had problems such as inter-frame interruption and frame omission due to telecommunication. Since the frames

The associate editor coordinating the review of this manuscript and approving it for publication was Ze Ji.

of the visible image and the IR image do not match, we tried to find the same frame between the two videos by finding a correlation through the image matching network. In the process, a study of matching networks was conducted, and this paper deals with the contents.

Visible/IR Cameras are passive sensors. The sensors' data measure different characteristics depending on the wavelength from the reflected spectrum of the natural or artistic illumination of the target to the emitted spectrum. Images acquired with various spectra have different characteristics depending on the wavelength of light. Visible sensors are using red, green, and blue (RGB). An Electro-Optical (EO) includes a wider spectrum than visible. So, EO/IR systems cover the range from ultraviolet (UV) through visible and IR. Each band has the different wave length (UV: 0.25∼0.38um, visible: 0.38∼0.75um, and infrared: 0.75∼14um). Visible, near IR(NIR), short wave IR(SWIR), and mid-wave IR (MWIR) measures reflected radiance, while long wave IR (LWIR) measures emitted radiation [1]. Therefore, judging various types of matching with heterogeneous sensors measuring different spectral bands as the same matching network may make an error of ignoring physical characteristics. We perform matching using visible and TIR data, which are aerial drone data using LWIR.

Our contribution to the matching of Visible and TIR can be summarized in three aspects.

–First, we compare both of matching network and classification network in matching scene.

–Second, we propose useful attention module for heterogeneous sensor matching.

–Third, we suggest suitable network for visible and TIR matching and execute the matching with directly acquired complex aerial drone data.

This paper explains the utility of the proposed edge attention by applying the attention block in heterogeneous sensor matching and proves that the stacked input can efficiently determine the matching. In addition, by using stacked input, we prove that the proposed network is robust through comparison with classification networks as well as matching networks. To learn TIR data, we used complex aerial drone data to prove the robustness of our network.

This paper shows the construction of a matching network through analysis of various parameters and present an efficient network to classify matching results of Visible and TIR. There are various cases of using stacked input for matching problems [2]–[4]. Since these stacked inputs can be reflected as a single 2-channel input or 4-channel input, it is required to compare between matching networks with classification networks. Additionally, Judging the matching of heterogeneous sensors as a classification network is a task that has not been done before.

From the point of view of heterogeneous matching, we tried to find the comprehensible module for alignment through the Attention Module, which is similar with human thinking. Figure 1 shows an example of registration and misregistration at the same time. Through what characteristics

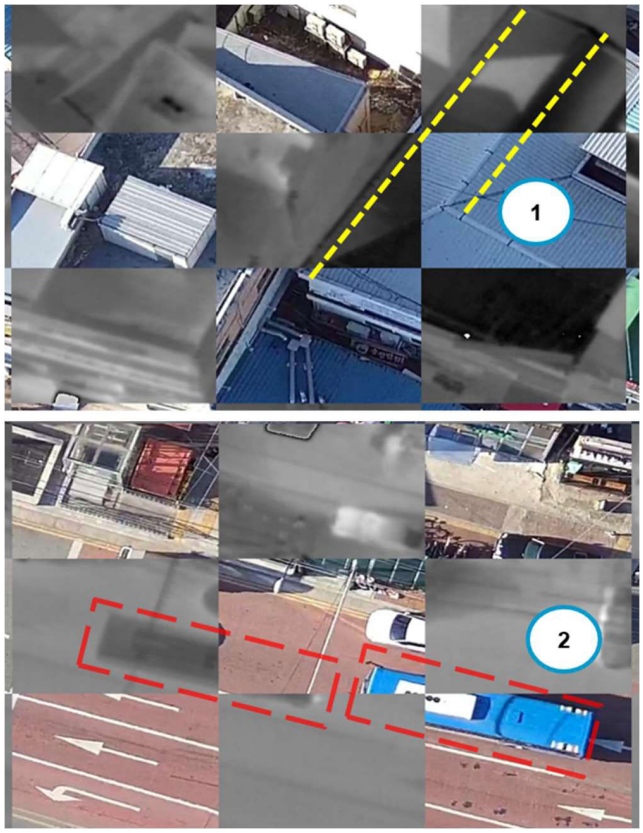

FIGURE 1. Examples of matching(up) and mismatching(down).

do you judge the right or wrong of matching in this figure? Our research team directly made the matching ground truth and judged that the primary focus of human visual judgment of matching between Visible and TIR is the aligned edge of Visible and TIR, and secondarily, it is texture. Yellow lines indicate that a matching of visible and TIR worked well. And Red lines show mismatch. Because bus parts of the under image do not match. The line information is important to distinguish image matching and the texture of the object is also necessary. Through this, we propose a module that works similarly to human thinking.

For the matching of Visible and TIR, we organized a section as follows: *2. Related Work, 3. Proposed Method, 4. Ablation study*, and *5. Conclusion*. In the *2. Related Work*, we explain various example of fusion network, and then focus on image sensors based heterogeneous matching. This section is divided into A) Heterogeneous Data Matching, B) Deep Learning for Classification, and C) Attention Module. In A) Heterogeneous Data Matching, examples of applying deep learning as well as conventional methods to matching heterogeneous data are described. We conducted a comparative experiment with the classifier in the experimental section, and contents related to the classifier were described in B) Deep Learning for Classification. C) Attention Module introduces the attention modules, a field that has recently been in the spotlight for Convolutional Neural Network (CNN) research.

3. Proposed Method explains the proposed network and its feasibility. AVIL block using attention module, AVILNet using the blocks, and loss function effective for binary classification are introduced in this section. *4. Ablation Study* proceeds with the acquisition procedure for experimental data and various network comparison experiments. At *5. Conclusion*, we conclude by explaining the effectiveness, practicality of this paper, and the future works.

II. RELATED WORK

This section consists of three parts. *A. Heterogeneous Data Matching* introduces various cases of deep learning used for image matching. *B. Deep Learning for classification* describes the history of deep learning classifier and introduces the convolution-based networks compared in this paper. Also, we explain why the classifier algorithm is used for matching heterogeneous sensors. *C. Attention Module* briefly explains the attention module and its usage examples. Before describing the sub-section, the fields of various sensors will be described.

Many sensors are being used in various fields. Lidar can acquire depth information of an object using point cloud and has the advantage of high precision, but it cannot obtain the color and surface characteristics of the object [5], [6]. X-rays used in medicine are produced when very fast-moving electrons collide with heavy atoms. Although there is an advantage of short-time examination, only rough information of soft tissue (subcutaneous-tissue /muscle/ligament) can be grasped using X-ray diffraction. Computed Tomography (CT) has the advantage of being able to check the cross-sectional view of an object, but it is also insufficient to measure soft tissues in detail. Such various single sensors have limitations, advantages, and disadvantages in the acquisition process and image information.

Multi-sensor fusion is being studied to solve and supplement the limitations of a single sensor. Fusion technologies used in the autonomous driving field are mainly Visible and Lidar convergence [5], [6]. Gong *et al.* [5] implemented fusion for 3D object detection using point cloud and visible information, and Caltagirone *et al.* [6] used deep learning-based fusion technology for path detection. In the medical field, fusion of heterogeneous sensors such as CT and X-ray was carried out. CT and Chest X-ray (CXR) fusion was implemented based on deep learning for Diagnosis of COVID-19 [7], and Panwar *et al.* [8] also suggested deep learning framework for detection of COVID-19 and proved the responsibility by using GradCAM [9]. These attempts are efforts to go beyond the limits of a single sensor.

Efforts to fuse heterogeneous sensors are no exception to fuse between heterogeneous images. There are many cases based on EO and IR fusion [2], [10]–[13]. A dual-tree complex wavelet transform (DTCWT) technique based on region segmentation was proposed for fusion of airborne infrared and visible images [10]. Fusion of low intensity visible and thermal infrared was performed, and frequency band fusion was performed after reinforcing low light information using IHS conversion [11]. Sensor fusion was performed using a CNN-based DeepFuse network and a learning loss based on the structural similarity index measure (SSIM) [12]. Li and Wu [13] proposed DenseFuse which is a dense block-based network and used a learning loss applied with SSIM. Sensor fusion results were acquired through deep learning.

For such fusion, matching between the two sensors must be performed, and fusion must be performed through correctly implemented matching. Using the reflection bands of 3 RGB channels and 1 NIR channel as inputs, Dense Block type network was designed and matching or mis-matching were determined [2]. Matching was performed on EO aerial data at different times using the Siamese network [14], and a 2ch network was suggested, which is a similarity measurement network between the visible band and near IR(NIR) [3]. Its input is Visible converted to gray level and NIR. The network consists of a total of three convolution layers. Zhang *et al.* [15] proposed a Siamese Network-based sFcNet. For EO/Near Infrared (NIR), EO/TIR, and EO/Synthetic Aperture Radar (SAR), first feature was acquired through convolution layers for high-resolution EO, respectively, and second feature was obtained through heterogeneous sensors. Using second feature as a filter for first feature, the matching score was determined. Wang *et al.* [16] proposed an algorithm that learns by vectorizing the patch of each image and finds the matching point. Through the following three sections, we introduce the related papers of the techniques used in this paper.

A. HETEROGENEOUS DATA MATCHING

For matching heterogeneous sensors, Scale Invariant Feature Transform (SIFT) or feature-based extraction methods have been mainly studied, but the similarity measurement using deep learning is currently being developed. Ma *et al.* [17] extracted matched pairs for aerial photographs using SIFT. In the process, they used a gradient magnitude of the Gaussian scale-space image by means of Sobel filters to create robustness of the descriptor. Ye *et al.* [18] measured the structural similarity between images and performed registration between EO, SAR and Lidar heterogeneous data using a histogram of orientated phase congruency (HOPC) descriptor. Li *et al.* [19] performed multi-modal image matching using Radiation Invariant Feature Transform (RIFT). RIFT uses phase congruency instead of image brightness to detect feature points, and extracts corner points and edge points for optical-optical, infrared-optical, Synthetic aperture radar (SAR)-optical, Map-optical, and day-night matching. There are examples of a combination of feature-based and area-based fusions [20], [21], and line feature-based fusion [22], [23].

Recently, a matching algorithm using deep learning is also being studied, and the framework of the study is shown in Figure 2. Figure 2(a) is a Siamese network, where input A and input B enter different inputs into the same network sharing weights [14], [24]–[29], and figure 2(b) is a case where the network structure is the same but does not share weights. Figure 2(c) shows that the stacked input enters a

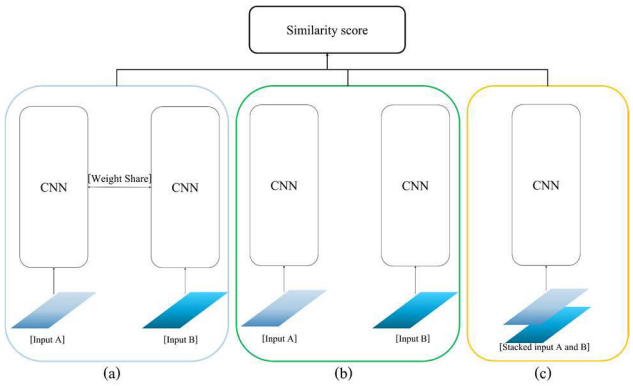

FIGURE 2. Framework of the matching networks [2]. (a) the siamese network, (b) the pseudo-siamese network, and (c) the channel-stacked input network.

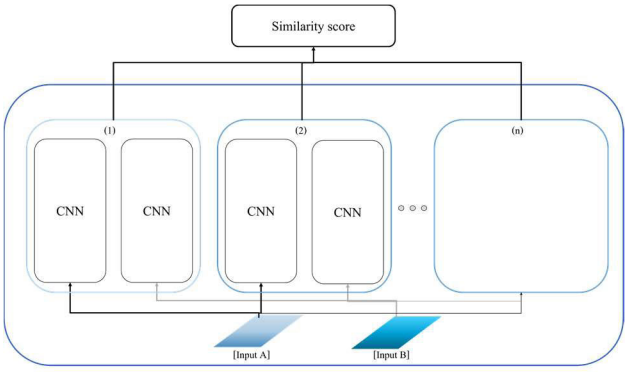

FIGURE 3. Framework of the n stream matching networks.

single network and measures the similarity [2], [4]. Zbontar and LeCun [24] performed stereo matching of the two visible data taken from the different angle, and He *et al.* [14] found a matching point for the EO data of the different weather and time zone. Han *et al.* [25] constructed a feature network and a metric network using MatchNet which considers various sizes of a patch. He *et al.* [26] proposed multi-support patches siamese networks (MSPSNs), and the registration was studied using satellite multispectral data (e.g., Landsat-5/8, ZY-3, and GF-1). Various sizes were reflected by adjusting the patches to sizes of 24 × 24, 48 × 48, and 97 × 97. There are also networks that have matches through various paths. Figure 3 is for an n-stream network. Suárez *et al.* [4] used the 2-stream network. En *et al.* [28] proposes three streams (TS-Net) and constructs a layer with two paths for a single input. It suggests two siamese networks to obtain three stream outputs. Balntas *et al.* [29] proposed a PN-Net (triplet network) and trained by receiving the same input pair w and x and a different input pair y as input. They used the 3-stream network. Aguilera *et al.* [27] also proposed a quadruplet network called Q-net, which uses two pairs of EO and NIR inputs and uses a total of four inputs.

Baruch and Keller [30] showed better performance on VEDAI, CUHK, and VIS-NIR data sets compared to Aguilera *et al.* [27], which is an example of a network that does not share network parameters in figure 2(b). They designed a block like figure 2(a) and 2(b) together and conducted research. Networks corresponding to figure 2(a)-(c) were made and tested [3], [4]. Aguilera *et al.* [3] proposed a 2ch network, which uses stacked input. Zagoruyko and Komodakis [31] further tested the 2 streams network and finally proved the robustness of 2 channel - 2 stream. Suarez *et al.* [4] designed a two-channel network similar with Aguilera *et al.* [3], but using fewer parameters, higher performance than Aguilera *et al.* [3] was derived in the match of visible and NIR. Higher performance than Aguilera *et al.* [3] was derived for visible and NIR aerial images using the dense block [2]. Vectors from feature are also performed for matching [16]. Chen *et al.* [32] presents FSNet which is kind of a siamese network and suggests registration for heterogeneous images. These developments in deep learning have made great strides in the performance of matching between two heterogeneous images. With reference to the history of this development, we conducted a study using deep learning. As the channel stack network is developed, we thought that it is necessary to consider the judgment of heterogeneous sensors matching through the developed deep learning classifier.

B. DEEP LEARNING FOR CLASSIFICATION

Recent research on deep learning has been inspired by the shape of the brain. Lenet5 [33] is a classic deep learning network used for text classification. The classification was performed on the 32×32 input using three convolution layers, two subsampling layers, and one fully connected layer. Deep learning networks have evolved in the direction of using a deep convolution layer while solving the problem of vanishing gradients through weight initialization [34], [35], batch normalization [36], etc. VGG [37] uses a 3×3 convolution to improve the classification performance for input data of different sizes through layers of various depths, and Resnet [38] improves the classification performance by using a residual block. ResNeXt [39] which use the grouping of filters using cardinality and the combination of the residual block used in Resnet [38] recorded high performance top-5 errors. Deep learning network has also developed into a form of accumulating channels, such as DenseNet [40]. In DenseNet, the performance of classification was improved by concatenating and using the previously used convolution block. Dual Path Network (DPN) [41] was designed as a model that utilizes both the advantages of ResNet and DenseNet using both residual and dense network paths for the dual path structure. MobileNet [42] simplifies networks by shortening them to fit mobile devices. Therefore, depthwise convolution and pointwise convolution were used to reduce the number of parameters and the amount of computation. ShuffleNet [43] uses point-wise group convolution and channel shuffle to create a small model to reduce the number of parameters and computational amount like MobileNet.

CSPNet [44] is an abbreviation of Cross Stage Partial Network, and the network's convolution layers were constructed

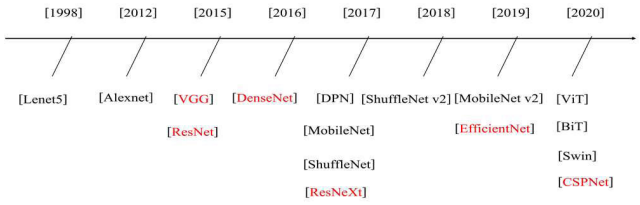

FIGURE 4. Simple history diagram of deep learning image classification network.

by dividing the base layer into parts and convolutional only part of it, and then merging the rest. EfficientNet [45], which is currently showing high performance in the public classification data Cifar10 and Cifar100 [46], improves performance through compound scaling in the direction of changing the size of various existing models such as width scaling, depth scaling, and resolution scaling. By using EfficientNet, the current best performance was derived from cifar10 and 100 through learning using the teacher network and the student network [47]. Recently, ViT [48], BiT [49], and Swin [50], which applied the transformer used in natural language processing to image classification, are also showing high performance in ImageNet [51]. However, as can be seen in figure 2(c), since the stacked data of heterogeneous images is used as an input, it is worth considering the image classification networks in the registration. Figure 4 is a schematic history diagram of the image classifiers which we explained. Networks written in red are used for comparison with proposal network in this paper. In this paper, we propose a matching network using CNN-based attention module. So, we compared the CNN-based networks highlighted in red at figure 4. The classification networks showed optimal performance in public data sets. Since the stacked 2 or 4 channels as input for classification network are proper, we judge that it could be sufficiently used for classification networks. Also, the optimal input of the network presented in this paper is 128×128, which is sufficient to apply to the classification network.

C. ATTENTION MODULE

We tried to improve the performance of the matching network by applying the attention module that improved the performance of the classifier. Therefore, in this section, the progress of the attention module will be briefly described. Attention modules improved classification performance [52]–[55]. SE block [52] which is an example of improving the performance by performing channel-oriented attention was designed. Channel attention was performed through global average pooling, and the channel was emphasized through a fully connect layer. A BAM block [53] used channel attention through global average pooling and spatial attention through 1×1 convolution. Woo *et al.* [54] presented a CBAM block, which uses features using max pooling as well as global average pooling additionally, and unlike the simultaneous usage of channel and spatial attention in BAM, first proceeds channel attention and additionally proceeds spatial attention. After that, apply residual blocks for reinforcing input features. A residual attention network was designed and applied to the classifier [55]. Channel attention was performed by designing RCAB [56], and the residual in residual concept was applied to image super-resolution. A channel-wise and spatial attention residual (CSAR) block was designed and used for super-resolution [57]. The advantage of channel attention is that it gives weights to important channels. However, there is a limit to not seeing spatial characteristics. To solve this spatial weakness, the attention block studies have been conducted in the direction of designing the spatial module, and the existing spatial attention block uses a square filter. This square filter is interpreted to mean that unnecessary spatial information is also emphasized in matching where edge information judgment is important. In this paper, we propose the validity of AVIR block emphasizing edge components through comparison with SE, BAM, and CBAM used in the classifier.

The next section is a description of the proposed network and loss, and additional settings are augmented through the experimental part.

III. PROPOSED NETWORK

The problem with the existing matching network is that it was not possible to acquire high matching rate for the TIR and Visible pairs of the drone aerial data. The data we used for research is drone data obtained directly, and there are various complex objects like buildings, trees, grass, and vehicles. TIR images are information containing the emitted radiance characteristics of various objects. So, the TIR images are different with visible images which are used reflection radiance. The data also contains many more complex features than distinct edge information. A dataset used at [58], [59] contains human radiance information and indoor-oriented data pairs, so the dataset has distinct characteristics that distinguish background and people information. It is necessary to discuss networks for complex aerial data.

In this paper, the proposed network using AVIR block is named AVIRNet. A 2-channel stacked input in which gray scale visible and 1-channel LWIR are concatenated is used as input. It was designed considering various single channel inputs (i.e. panchromatic, LWIR, SWIR, and MWIR images) used in aerial field. It has very concise layers, a total of 5 convolution blocks, 5 AVIR blocks, and 5 pooling to form a layer, and the features after the last 2d convolution are input to the fully connected layer through global average pooling. The 5 convolution blocks used for this include 3×3 filter, zero padding of 1, stride of 1 interval, batch normalization [36], and ReLU. Finally, only a scalar value is reduced to 0 and 1 through the sigmoid function, 0 means mismatching and 1 means matching. Figure 5 is a schematic structure of the network, and Table 1 is a description of each block of AVIRNet. CB is convolution block which consists of convolution of 3×3 filter, batch normalization, and ReLU. AT stand for attention module. MP stand for a max pooling. GP stand for a

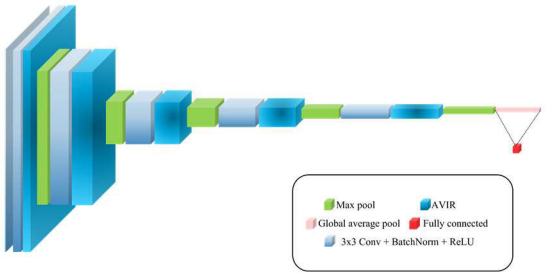

FIGURE 5. The architecture of the AVIRNet.

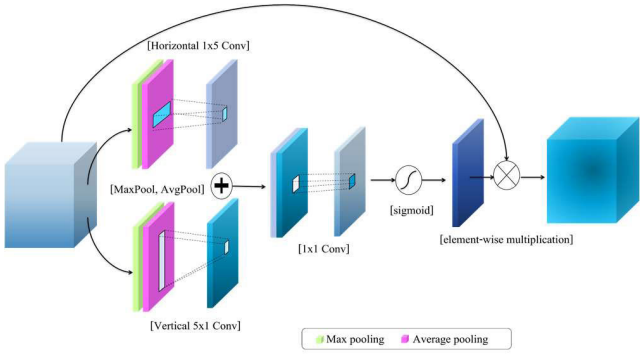

FIGURE 6. The AVIR block.

global average pooling. FC is a fully connected layer. Filters of AT are composed of three types, which are horizontal, vertical, and spatial filters in order. P stands for padding and S stands for stride. Figure 6 is the AVIR block. Max pooling and average pooling make different feature from previous feature. After Horizontal and Vertical convolution, feature can concentrate in an edge information. 1×1 convolution makes combination of the horizontal and vertical edge information. Sigmoid function can compress the result.

The following is the description of the AVIR block. When designing the attention map, the most focused information was judged to be edge information through manual registration. To this end, we designed an edge module through filters on the horizontal and vertical axes to include less spatial information. Assuming F as a feature after convolution, channel-wise max pooling and avg pooling for the feature are equal to $F_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F_{max} \in \mathbb{R}^{1\times H\times W}$. Each pooling is indicated by $F_{avg} = AvgPool\,(F)$ and $F_{max} = MaxPool\,(F)$, and the edge module can be expressed as a horizontal convolution module and a vertical convolution module as in (1) and (2).

$$E_x\,(F) = f^{1\times s_x}[F_{avg}; F_{max}] \tag{1}$$
$$E_y\,(F) = f^{s_y\times 1}[F_{avg}; F_{max}] \tag{2}$$

Equation (1) is designed so that horizontal information can be judged deeply by configuring a horizontally long filter, and (2) is configured with a vertically long filter. So, each feature makes it easy to utilize horizontal and vertical information. As for the padding used for each filter, the quotient of filter size divided by 2 is applied to the horizontal and vertical axes on both sides of the feature. [A; B] means concatenate between A and B. Equation (1) is a $1 \times s_x$ filter designed in the direction of the x-axis, and the left of Figure 7 is an example of a convolution diagram using a 1×5 filter. Also, (2) is a filter designed in the y-axis direction, as shown in the middle of Figure 7. In addition, the combined feature through (3) is extracted through spatial convolution of $s_a \times s_a$ size of horizontal information and vertical information, and the final value of each pixel is compressed to 0-1 through the sigmoid function marked σ in (4). The feature of the final attention model is the dimension of $M(F) \in \mathbb{R}^{1\times H\times W}$. The strength of this Edge Module is to expand the judgment of edge information to overall spatial information by judging horizontal and vertical information as different paths. As we change the values of s_x, s_y, and s_a we deal with the filter size most appropriate for the matching rate of aerial data through an ablation study.

$$S\,(F) = f^{s_a\times s_a}[E_x\,(F)\,; E_y\,(F)] \tag{3}$$
$$M\,(F) = \sigma(S(F)) \tag{4}$$

The learning loss used a binary cross entropy loss for finding the correct answer and a smoothing term. Human can judge matching as 1 and mismatching as 0, but in machine learning, the concept of probability distribution, there cannot be a perfect integer result. Therefore, the smoothing term was

TABLE 1. Network parameters of avirnet.

Name	Type	Output dim.	Filter size	P	S
Conv1	CB	128×128×64	3×3	1	1
AVIR	AT	128×128×64	1×5, 5×1, 1×1	2,2,0	1
Pool	MP	64×64×64	2x2		2
Conv2	CB	64×64×128	3×3	1	1
AVIR	AT	64×64×128	1×5, 5×1, 1×1	2,2,0	1
Pool	MP	32×32×128	2x2		2
Conv3	CB	32×32×256	3×3	1	1
AVIR	AT	32×32×256	1×5, 5×1, 1×1	2,2,0	1
Pool	MP	16×16×256	2x2		2
Conv4	CB	16×16×512	3×3	1	1
AVIR	AT	16×16×512	1×5, 5×1, 1×1	2,2,0	1
Pool	MP	8×8×512	2x2		2
Conv5	CB	8×8×512	3×3	1	1
AVIR	AT	8×8×512	1×5, 5×1, 1×1	2,2,0	1
Pool	MP	4×4×512	2x2		2
pool	GP	512	4×4		1
FC	FC	1	512		

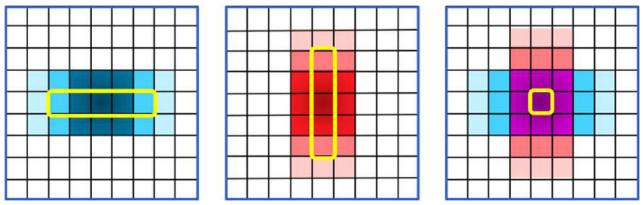

FIGURE 7. Horizontal convolution filter(left), vertical convolution filter(middle), a feature after the concatenate process(right).

FIGURE 8. The DJI M200 drone and DJI XT2 camera.

additionally set to make it impossible to have perfect 0 and 1, and the performance improvement is also presented through an ablation study that is higher than the case of using a single binary cross entropy for finding the correct answer.

$$L = -y_i \, log \, (N \, (x_i; \theta)) - (1 - y_i) \, log \, (1 - N \, (x_i; \theta)) \tag{5}$$

$$L' = (1 - \epsilon) \, L \, (y_i, N \, (x_i; \theta)) + \epsilon L \left(\frac{1}{2}, N \, (x_i; \theta) \right) \tag{6}$$

L of (5) is an abbreviation for binary cross entropy loss. y_i is a ground truth and N (x_i; θ) is the output of the input $\boldsymbol{x_i}$ by the model parameter θ. Equation (6) is a loss equation which contains smoothing term used in the experiment. For the loss ratio, the ratio of binary cross entropy loss and smoothing term is determined by the ϵ value, and 0.05 was used in the experiment of this paper. $N \, (x_i; \theta)$ is the probability result and is the value after the sigmoid function. y_i is the desired learning target, meaning 0 for mismatching and 1 for matching. $L \, (1/2, p)$ is the smoothing term from the center of [0, 1] when $p = N(x_i; \theta)$. This smoothing term helps the resulting values converge slightly by 0.5.

IV. EXPERIMENT

A. DATASET

Drones with visible/TIR cameras are one of the newest devices that can be organically used in civilian and defense applications to monitor objects day and night. We acquired drone data for surveillance and reconnaissance research and confirmed that the two images have different fps due to different devices. To solve this problem, a matching network was implemented to check the correlation between two images. For the visible/TIR data set, a matching set was constructed using the aerial data measured with DJI M200 model drone and DJI XT2 camera. Through this, four locations in Gyeongsan, Gyeongsangbuk-do, Republic of Korea were filmed. The XT2 can shoot both a thermal camera and a visual camera at the same time, and the spectral range for the thermal imaging camera is 7.5 to 13.5$\boldsymbol{\mu}$**m**. The possible shooting temperature includes the range of −25 to 135°C. The spatial resolution is 640 × 512 and it has an operating cycle of 30 fps. The visual camera has a FOV of 57.12°×42.44° and a maximum resolution of 3840 × 2160. Figure 8 shows the DJI M200 drone and DJI XT2 camera.

FIGURE 9. Thermal infrared image(left) and visible image(right).

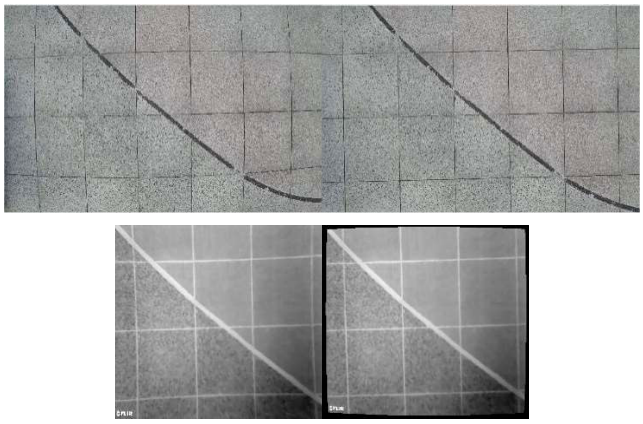

FIGURE 10. Distortion calibration images of visible image(up) and thermal infrared image(down).

Figure 9 is examples of an aerial image. The following is a description of the overall process for the acquired drone dataset. The drone data of the EO/IR pair we acquired was taken from the same position and angle. Due to the characteristics of the drone, when we check the acquired images, the Barrel distortion effect in EO images and the pincushion distortion effect in TIR images are occurred. To geometrically correct it, we measured internal and external parameters using Matlab and geometric correction was performed using the camera calibration toolbox [60]. Also, registration between two images was performed through homography registration.

Among Matlab toolbox, Camera Calibrator can check internal and external parameters of the camera and identify lens distortion. It includes a camera calibration function. In addition, calibration should be performed using the chess board image acquired using the Visible/IR camera. In the case of aerial photography, since the focal point is at a far distance, correcting the chess board image measured at a short distance may cause a large error for long distance image registration. Thus, a black-and-white image of a chess board shape as shown in figure 10 was obtained using a 60cm x 60cm tile sculpture measured from above, and the obtained internal parameters are shown in figure 11.

The Figure 10 was obtained after correcting for lens distortion. The transformation between two coordinates by homography is shown as (7). Since the two sensors acquire images using a single drone, by obtaining the image matching parameters only once, all photos at similar height can use the single

	1	2	3
1	3.4650e+03	0	0
2	0	3.5031e+03	0
3	330.0458	221.2014	1

	1	2	3
1	3.9720e+04	0	0
2	0	3.9249e+04	0
3	1.9205e+03	1.0074e+03	1

FIGURE 11. Cameras intrinsic parameter of visible camera (left) and TIR camera (right).

homography matrix. In coordinate system transformation using homography, when four matching pairs are obtained for two matching points p and p', coordinate transformation is possible through the following matrix in (7). In the formula, if the coordinates corresponding to x, y, and 1 on the right are p, the coordinates corresponding to wx', wy', and w are p' [61].

$$\begin{bmatrix} wx' \\ wy' \\ w \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{7}$$

Finally, Figure 12 was obtained through lens distortion correction and homography correction to match visible and TIR. The problem with data generated in this way is that data generated by manual matching of heterogeneous sensors has pixel errors. The first cause is an error caused by geometric correction. The second cause is the error due to the distance between the stereo cameras. The minimum pixel error of the data we finally computed can yield an error of 0 to 5 pixels, and since the error can prevent the network learning, we scale the original data by 1/4 to reduce the error to 0 to 1.25 pixels. In addition, due to the limitation of manual homography stereo matching, the error widened toward the outside, so the experiment was conducted using 1024×1024 in the center of the image. In various papers, there have been cases in which data sets are constructed with an input size of 64×64 and a smaller input size of 36×36. However, considering the nature of the aerial data, the input size was selected as 128×128 in consideration of the data information. Too small image patch does not have enough information for registration. Through preprocessing, we obtain matching patches at Figure 13. We filmed a variety of environments, including farmland, settlements, rivers, college towns, roads, and forests. Scene 1 is the university interior and forest, scene 2 is the driveway and building, scene 3 is the village, forest, parking lot and driveway, and scene 4 is around the university's main gate, farmland, and stream. Since the data cannot be disclosed due to internal security issues, a clear designation is omitted. The data structure was divided into 4 data taken at different locations as shown in Table 2. Training data consist of scenes 1, 2, and 3 and validation data consist of scenes 3 which is different of training data but has similar aerial property with training data. In addition, test was conducted using the parameters of the epoch with the lowest validation loss, and the test was configured using scene 4. Due to the nature of the aerial drone dataset, we often observe unlearned landscapes, so it was judged that it was the right experimental result to obtain a robust network even in scene 4 that is not related to the train set.

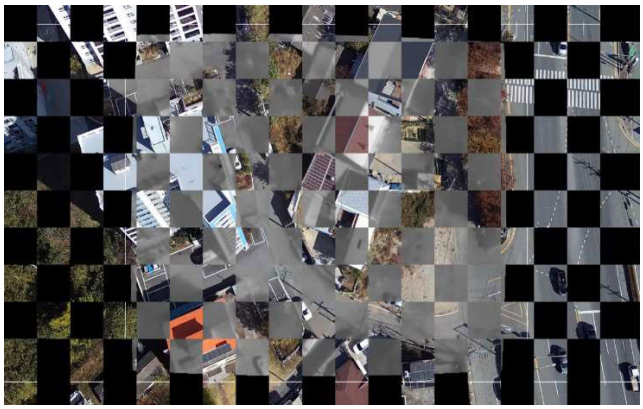

FIGURE 12. Visible and thermal-IR matching result.

TABLE 2. Train, validation, and test set.

data	scene	The number of pictures (size)	The number of matching patches
Train	Scene 1 Scene 2 Scene 3	373 (258x258)	3,357
Val	Scene 3	124 (258x258)	1,116
Test	Scene 4	123 (258x258)	1,107

B. EXPERIMENT SETTING

For experiments, AdamW [62] was used, learning rate was 0.001, and the beta values were 0.9 and 0.999. The learning rate was adjusted using the cosine annealing scheduler [63]. The training epoch used is 100 epochs. Batch size is 16 for train. In addition, the data loader stage was configured as shown in figure 14 to learn various mismatching for learning. It is judged to be 1 in the case of a matching patch and 0 in the case of mismatching. In addition, if the r distance is set in the data loader, the data loader is configured so that learning can be robust even at various pixel distances by using mismatching patches at random angles for data that deviate by the corresponding distance from the center of the reference image. The configuration of this data loader can increase the diversity for mismatching pairs, and the range from 0 to 360° is used, and r uses the range of [50, 70] to increase the diversity through a random distribution in which the input is not determined. A random flag was given with a probability of 1/2 during training so that mismatching and matching data automatically occurred. Figure 14 shows the data loader.

C. ABLATION STUDY

In this paper, a total of 11 network experiments were conducted, and the networks used were 5 matching networks and 6 classification networks. We use overall accuracy (OA) of (11) and mean accuracy (MA) of (12). Because A random flag with a fixed seed is used in data loader. TP means that a classification result is positive, and the result is correct. FN means that a classification result is negative, and the result

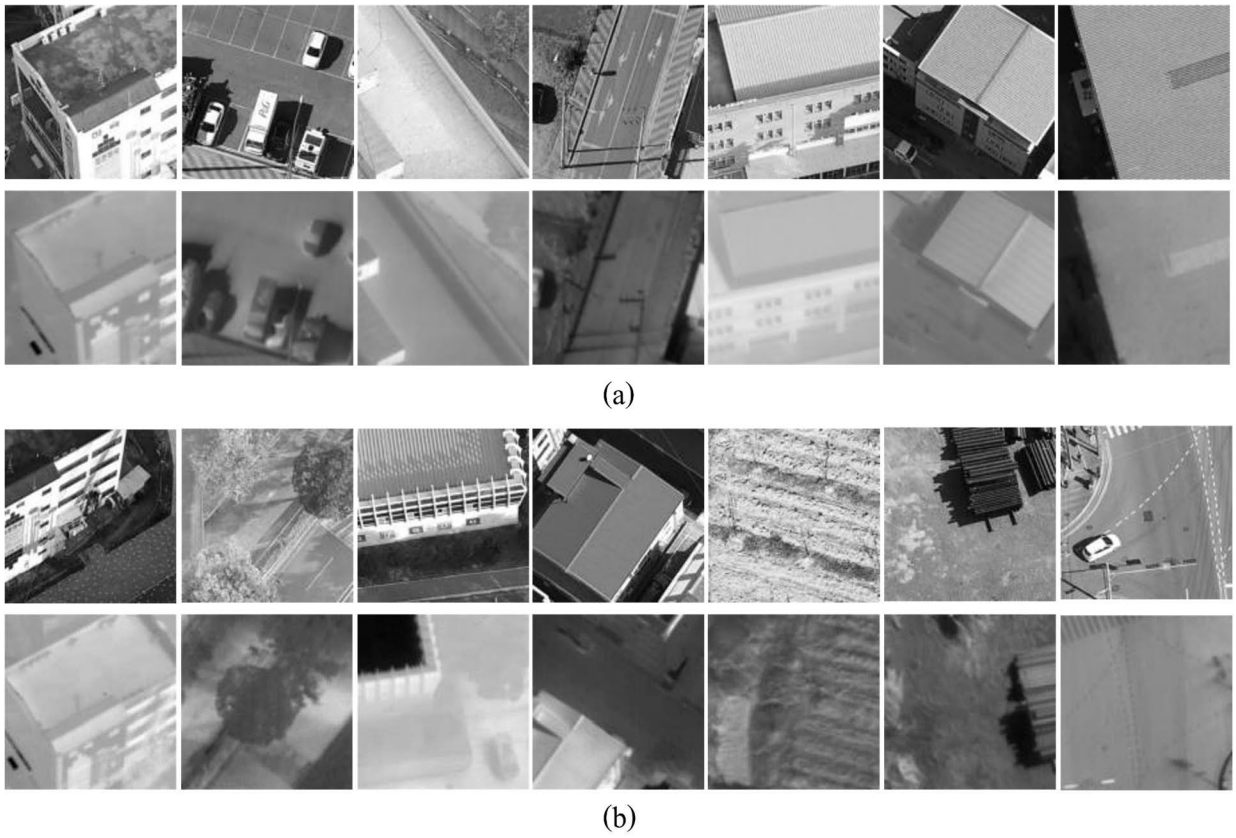

FIGURE 13. Data examples. (a) matching images, (b) mismatching images.

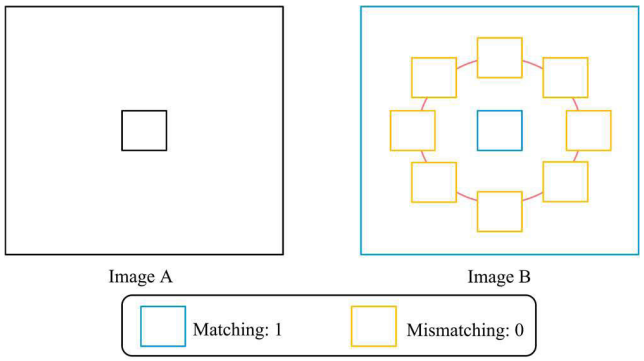

FIGURE 14. Matching data loader.

is incorrect. TN means that a classification result is negative, and the result is correct. We use true positive rate (TPR) of (8) and true negative rate (TNR) of (9) for calculating accuracy.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$TNR = \frac{TN}{TN + FP} \tag{9}$$

$$MA = \frac{TPR + TNR}{2} \times 100(\%) \tag{10}$$

$$OA = \frac{\sum (TP + TN)}{All} \times 100(\%) \tag{11}$$

Table 3 is the result according to the operation of the AVIR block. (*) means a broadcast element-wise multiplication. We think that broadcast element-wise multiplication had a direct effect on features and (+) was done to avoid excessive data loss. This is because, when the AVIR block is applied directly to the broadcast element-wise multiplication to the feature, most values except for the edge become close to 0. So, (+) can prevent values from disappearing. In this experiment, only broadcast element-wise multiplication applied gave the best performance.

we used batch normalization, and through this, learning was carried out through edge emphasis, which was appropriate to keep the feature values from being completely zero. Figure 15 shows features at each stage of AVIR block. The vertical features are from (2) and the horizontal features are from (1) at every attention block. We denote area which is from (4). When looking at these features intuitively, it can be confirmed that edge information is detected in the first AVIR block. When the filter is designed long, the network learns edge information by itself. As the network deepens, various convolution layers are applied, and the shape of the feature extracted from the AVIR block is transformed. We extract the features from every attention block and each feature size is

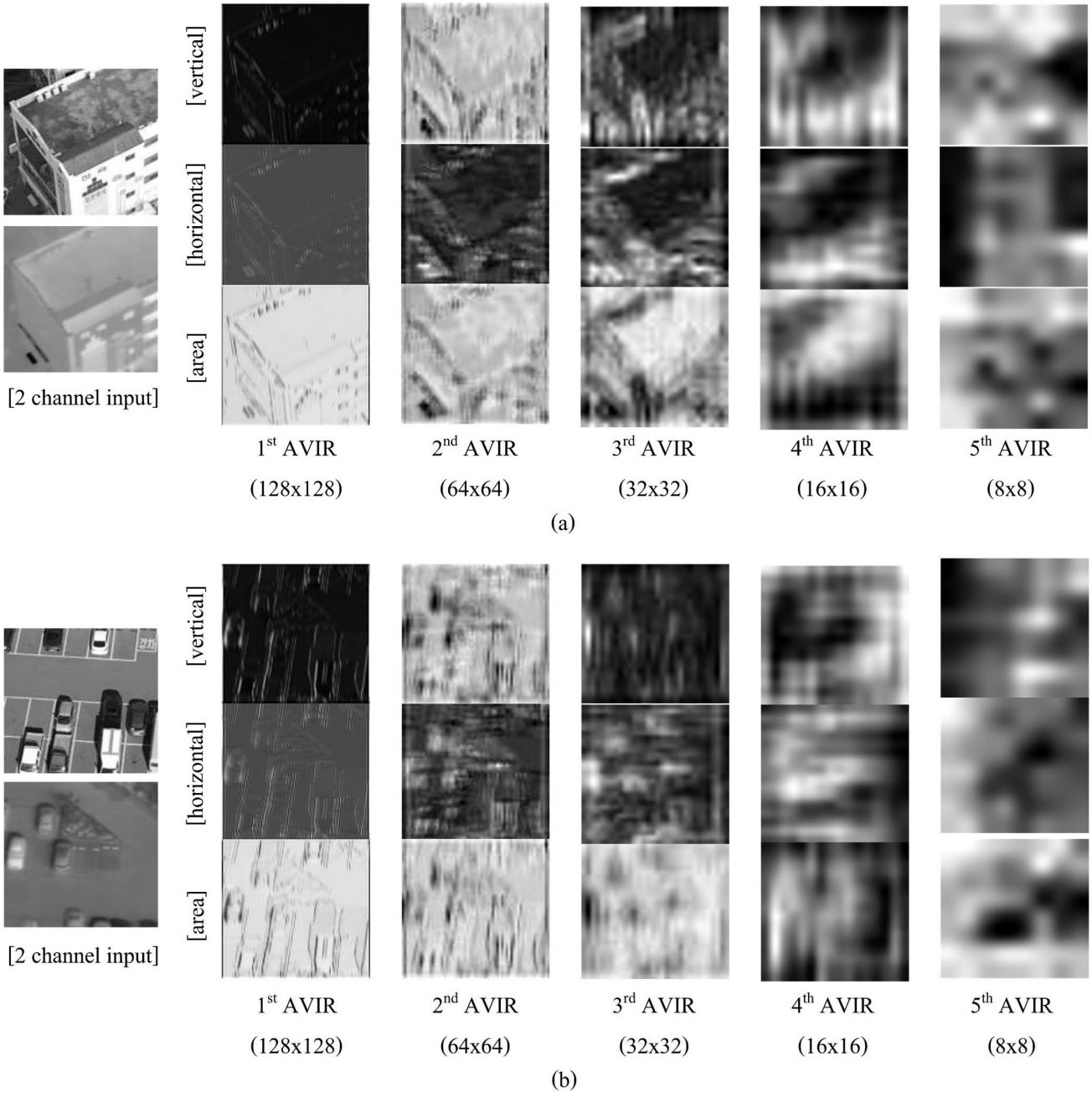

FIGURE 15. AVIR Block features. (a) is matching case and (b) is non-matching case.

128×128, 64×64, 32×32, 16×16, and 8×8 from the left. After the first AVIR block the result show the edge emphasis image of visible and TIR. The network can automatically learn the shape of edge and we prove the edge through visualization. Table 4 is the comparison of the result according to the loss. The smoothing term of the proposed loss was found to have a large effect. From a machine learning point of view, the results of 1 and 0, which are complete integer labels, do not exist probabilistically. Therefore, using the smoothing term resulted in good performance.

Table 5 describes the results according to the filter size of the AVIR block, and heuristically, the highest matching rate was obtained for s_y, s_x at 5 and s_a at 1. Using AVIR block shows better performance than not using AVIR block. In the test, when the AVIL block was not used and the AVIL block of s_y, s_x at 5 and s_a at 1 was compared, the OA increased by about 8.401%.

TABLE 3. Comparison of operator at AVIR block.

Filter	Val		Test	
	OA	MA	OA	MA
M(F) + F	92.025	92.049	91.509	91.494
M(F) *F + F	98.566	98.522	92.683	92.568
M(F)*F	**99.283**	**99.288**	**97.019**	**96.976**

TABLE 4. Comparison of loss.

Filter	Val		Test	
	OA	MA	OA	MA
Cross Entropy Loss	97.491	97.482	91.599	91.474
Proposed Loss	**99.283**	**99.288**	**97.019**	**96.976**

TABLE 5. Comparison of AVIR block's filter size.

Filter $(s_y, s_x), s_a$	Val		Test	
	OA	MA	OA	MA
Without	97.581	97.538	88.618	88.428
3, 1	96.595	96.505	94.490	94.403
3, 3	98.806	98.806	94.761	94.678
3, 5	98.656	98.657	93.135	93.024
3, 7	98.029	98.051	95.845	95.788
5, 1	**99.283**	**99.288**	**97.019**	**96.976**
5, 3	99.283	99.282	93.767	93.677
5, 5	98.595	98.595	95.664	95.613
5, 7	97.312	97.272	91.238	91.100
7, 1	98.835	98.837	94.038	93.943
7, 3	97.043	97.055	89.702	89.528
7, 5	99.194	99.214	96.296	96.238
7, 7	95.609	95.534	89.070	88.885

Table 6 is a comparison experiment between matching networks. (×) mark in the table means that the network could not learn the data. As a result of the experiment, AVIRNet obtained the dominant result. In the experimental process, Dense-based network [2], 2ch-2stream [31], and TS-Net [28] all performed matching learning with visible and NIR, the reflection band, and 2ch [3] and Domain Siamese [64] performed visible and TIR. Although we conduct same setting experiment with domain Siamese network [64], Learning did not proceed. The boundary of the used data is clearer, and the characteristics are clearer than our data, because the experiment was conducted using [58], [59], which is data with a clear temperature difference in an indoor space. In addition, there is a difference in information between the information around people of TIR used in the experimental process and the aerial data used in this paper. An experiment was conducted using the data of the size used in each paper. On the other hand, none of the networks learned. Furthermore, although the networks learned using the data size 128 × 128, they did not learn when the data was resized to the size used in each paper.

Since the input was 128 × 128, we could proceed with the experiment using the classifier network. To experiment with classifiers for different input sizes, we removed the flatten function before the fully connected layer of all networks and the multiple fully connected layers after convolutional layers. We replaced fully connected layers with a single fully connected layer using global average pooling. This was changed because the input data size of different networks was not constant, and it was not learned in the experiment using the flatten function of the fully connected layer used in the existing matching network [2], [3], [31]. In VGG 5, 6, and 7,

TABLE 6. Comparison of registration networks.

Model	Val		Test	
	OA	MA	OA	MA
Dense [2]	×	×	×	×
2ch [3]	×	×	×	×
2ch-2stream[31]	×	×	×	×
Domain Siamese[64]	×	×	×	×
TS-Net[28]	77.688	77.673	76.784	76.810
AVIRnet (proposed)	**99.283**	**99.288**	**97.019**	**96.976**

our research team confirmed the result of not learning in the network study of a single fully connected layer using flatten, so we obtained the results shown in Table 7 by using global average pooling instead of vector flatten. In addition, it was confirmed that learning of 2ch-2stream [31] was performed by changing the fc layer using flatten to the fc layer using global average pooling and changing the convolution filter size from 5×5 to 3×3. Because vector flatten ignores channel information and spatial information, it is not considered to be suitable for learning complex data and large-size images. VGG includes 3 fc layers in the network itself, but it is removed and replaced with 1 fc layer using global average pooling.

We also constructed the network in Table 7 to test how the change in the number of convolutions in the existing classification network affects. VGG 5, 6, and 7 do not exist at official paper. These use 2, 3, and 4 of the first convolution blocks (convolution + batch normalization + ReLU) of VGG9 and 2, 3, and 3 max pooling, respectively. ResNeXt also, ResNeXt11, 20, 38, 47, 56 layers do not exist as official. This was expressed as 11, 20, 29, 38, 47, and 56 for the case of using 1 to 6 Residual Blocks of ResNeXt29, respectively. As a result of the experiment, it was confirmed that many flops and parameters does not obtain good results in matching unlike classifiers, and the performances of VGG19, ResNeXt38, and DenseNet121 are good in the existing network. However, since it is necessary to find a matching point in a large image, the number of network parameters and the number of floating point operations (FLOPs) have a large effect on the time to find a matching point. AVIRNet is 1.792% higher in validation OA and 1.897% higher in test OA than VGG19 which is second highest. The multiplier-accumulate (Mac) is 4.89 GMac for VGG19 and 0.83 GMac for AVIRNet, which is ×5.89 difference. As the number of parameters, VGG19 is 20.04M and the proposed network is 3.91M, which is ×5.12 difference. These results show that AVIRNet using horizontal and vertical filters effectively finds features compared to existing networks using 3 × 3 convolution as a basic filter.

The following explains the robustness of the edge module in matching through comparison with the existing attention module. The attention blocks used for comparison are SE [52], which is channel attention, BAM [53] and

TABLE 7. Comparison of classification networks.

Model	Val		Test	
	OA	MA	OA	MA
VGG5	87.097	87.149	71.454	71.586
VGG6	95.699	95.773	87.805	87.744
VGG7	95.161	95.113	89.521	89.431
VGG9	98.477	98.460	91.328	91.192
VGG11	96.685	96.598	88.889	88.701
VGG16	98.297	98.329	93.677	93.588
VGG19	97.491	97.464	95.122	95.074
Resnet18	94.803	94.681	80.668	80.359
Resnet34	98.566	98.522	80.036	79.688
Resnet50	96.147	96.011	87.986	87.776
Resnet101	96.416	96.314	83.198	82.911
ResNeXt11 (32x4d)	91.577	91.726	79.404	79.506
ResNeXt20 (32x4d)	97.401	97.377	93.135	93.030
ResNeXt29 (32x4d)	98.387	98.434	93.677	93.606
ResNeXt38 (32x4d)	98.566	98.522	94.941	94.862
ResNeXt47 (32x4d)	98.746	98.720	89.160	88.971
ResNeXt56 (32x4d)	95.699	95.566	87.895	87.696
ResNeXt50 (32x4d)	99.194	99.177	89.612	89.443
ResNeXt101 (32x4d)	99.194	99.177	93.857	93.756
DenseNet121	97.670	97.631	95.032	94.960
DenseNet169	95.520	95.374	88.347	88.143
DenseNet201	94.982	94.927	87.986	87.776
CSP Resnet50	93.817	93.813	85.727	85.490
CSP Resnet101	95.789	95.860	88.799	88.687
CSP Resnet152	96.326	96.227	81.391	81.082
EfficientNet (b0)	×	×	×	×
EfficientNet (b1)	×	×	×	×
EfficientNet (b2)	×	×	×	×
EfficientNet (b3)	×	×	×	×
EfficientNet (b4)	×	×	×	×
AVIRnet (proposed)	**99.283**	**99.288**	**97.019**	**96.976**

TABLE 8. Comparison of attention module.

Attention Module	Val		Test	
	OA	MA	OA	MA
SE [52]	95.430	95.318	90.786	90.631
BAM [53]	93.369	93.209	90.515	90.349
CBAM [54]	95.341	95.188	89.341	89.161
AVIR (proposed)	**99.283**	**99.288**	**97.019**	**96.976**

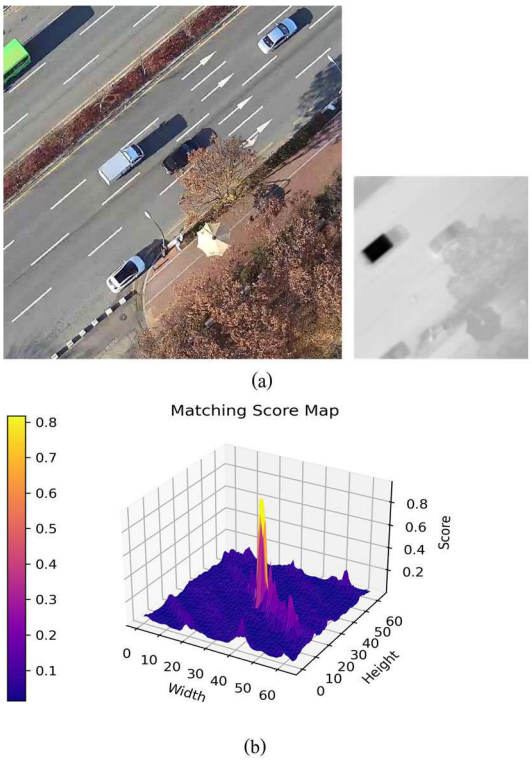

FIGURE 16. Matching Graph. (a) a visible image and a center 128 × 128 image of the TIR, (b) matching score map.

CBAM [54], which use channel attention and spatial attention at the same time. These were tested by substituting the AVIR block of AVIRNet, and as a result, it was confirmed that the AVIR block obtained a high matching rate. This result indicates that emphasizing edge information is more efficient in matching than emphasizing spatial information.

Figure 16 shows the matching score map of sliding window from a test dataset. We execute sliding window for visualization of matching score map. The interval of sliding window is 2 pixel and the result show the highest value at center point. We do not recommend to learning too short r distance at data loader. A very small r is considered non-matching, and when the label is set to 0, it may actually be positive due to an image pixel error. the resulting value can be larger than the threshold value of 0.5. In conclusion, the highest value was obtained at the center point (128, 128) in this example.

V. CONCLUSION

The use of large size input can handle registration information through feature learning rather than edge information. Therefore, the matching result, which was strong in the validation set, can lead to a low result in the test set that was not used for learning, and these results may cause unexpected problems in the problem of automatic registration. We applied the edge attention module to construct a network that can derive robust matching results even for unlearned data, although the input is larger than that of the existing matching networks. In addition, we proposed a matching network suitable for flight data matching and obtained 1.897% higher

matching overall accuracy even when 11 convolution layers were insufficient than VGG19, which performed the best in the existing classification network. Efficient removal of network parameters and judgment of matching results over existing layers through the attention module are meaningful because they are like human visual effects. As a result, it showed 6.233% higher performance than the SE [52] block, which showed the best performance among attention modules. we hope not only the effect of increasing the matching rate due to the addition of convolution in various framework configurations of the matching network, but also the effect of increasing matching rate using the attention module. This process is expected that it makes easy to design a fusion model for detection through preprocessing of EO/IR data that finds matching points. Recently, it has been found that the performance of object classification and detection using a transformer is excellent. This shows better performance than CNN-based technology when learning through a lot of data. With the advancement of these technologies, the matching of heterogeneous images should also be studied in the direction of deriving high accuracy by learning a lot of data. This is because, in the case of night, it is difficult to detect an edge compared to the daytime and scattering by light is sufficient to prevent common edge detection between heterogeneous images. So, deriving more precise results through learning a lot of data will show excellent performance in day and night surveillance and reconnaissance. Our research team plans to conduct research using transformers in the future.

REFERENCES

[1] G. Koretsky, J. Nicoll, and M. Taylor, "A tutorial on electro-optical/infrared (EO/IR) theory and systems," Inst. Defense Analyses, Alexandria, VA, USA, Tech. Rep. DTIC Document D-4642, 2013.

[2] R. Zhu, D. Yu, S. Ji, and M. Lu, "Matching RGB and infrared remote sensing images with densely-connected convolutional neural networks," *Remote Sens.*, vol. 11, no. 23, p. 2836, Nov. 2019, doi: 10.3390/rs11232836.

[3] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 267–275, doi: 10.1109/CVPRW.2016.40.

[4] P. L. Suarez, A. D. Sappa, and B. X. Vintimilla, "Cross-spectral image patch similarity using convolutional neural network," in *Proc. IEEE Int. Workshop Electron., Control, Meas., Signals Appl. Mechatronics (ECMSM)*, May 2017, pp. 1–5, doi: 10.1109/ECMSM.2017.7945888.

[5] Z. Gong, H. Lin, D. Zhang, Z. Luo, J. Zelek, Y. Chen, A. Nurunnabi, C. Wang, and J. Li, "A frustum-based probabilistic framework for 3D object detection by fusion of LiDAR and camera data," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 90–100, Jan. 2020.

[6] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "LIDAR–camera fusion for road detection using fully convolutional neural networks," *Robot. Auton. Syst.*, vol. 111, pp. 125–131, Jan. 2019.

[7] D. Sharifrazi, R. Alizadehsani, M. Roshanzamir, J. H. Joloudari, A. Shoeibi, M. Jafari, S. Hussain, Z. A. Sani, F. Hasanzadeh, F. Khozeimeh, A. Khosravi, S. Nahavandi, M. Panahiazar, A. Zare, S. M. S. Islam, and U. R. Acharya, "Fusion of convolution neural network, support vector machine and Sobel filter for accurate detection of COVID-19 patients using X-ray images," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102622, doi: 10.1016/j.bspc.2021.102622.

[8] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110190, doi: 10.1016/j.chaos.2020.110190.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[10] Y. Zuo, J. Liu, G. Bai, X. Wang, and M. Sun, "Airborne infrared and visible image fusion combined with region segmentation," *Sensors*, vol. 17, no. 5, p. 1127, May 2017, doi: 10.3390/s17051127.

[11] S. Liu, Y. Piao, and M. Tahir, "Research on fusion technology based on low-light visible image and infrared image," *Opt. Eng.*, vol. 55, no. 12, Dec. 2016, Art. no. 123104, doi: 10.1117/1.OE.55.12.123104.

[12] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732.

[13] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[14] H. He, M. Chen, T. Chen, and D. Li, "Matching of remote sensing images with complex background variations via Siamese convolutional neural network," *Remote Sens.*, vol. 10, no. 3, p. 355, Feb. 2018.

[15] H. Zhang, W. Ni, W. Yan, D. Xiang, J. Wu, X. Yang, and H. Bian, "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3028–3042, Aug. 2019, doi: 10.1109/JSTARS.2019.2916560.

[16] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 148–164, Nov. 2018.

[17] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, and L. Liu, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[18] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, Mar. 2017.

[19] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020, doi: 10.1109/TIP.2019.2959244.

[20] Y. S. Kim, J. H. Lee, and J. B. Ra, "Multi-sensor image registration based on intensity and edge orientation information," *Pattern Recognit.*, vol. 41, no. 11, pp. 3356–3365, 2008.

[21] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014, doi: 10.1109/TGRS.2013.2281391.

[22] C. Zhao and A. A. Goshtasby, "Registration of multitemporal aerial optical images using line features," *ISPRS J. Photogram. Remote Sens.*, vol. 117, pp. 149–160, Jul. 2016.

[23] O. Arandjelović, D.-S. Pham, and S. Venkatesh, "Efficient and accurate set-based registration of time-separated aerial images," *Pattern Recognit.*, vol. 48, no. 11, pp. 3466–3476, Nov. 2015.

[24] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1592–1599.

[25] X. F. Han, T. Leung, Y. Q. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3279–3286.

[26] H. He, M. Chen, T. Chen, D. Li, and P. Cheng, "Learning to match multitemporal optical satellite images using multi-support-patches Siamese networks," *Remote Sens. Lett.*, vol. 10, no. 6, pp. 516–525, Feb. 2019.

[27] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *Sensors*, vol. 17, no. 4, p. 873, 2017.

[28] S. En, A. Lechervy, and F. Jurie, "TS-NET: Combining modality specific and common features for multimodal patch matching," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 3024–3028.

[29] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: Conjoined triple deep network for learning local image descriptors," 2016, *arXiv:1601.05030*.

[30] E. Ben Baruch and Y. Keller, "Joint detection and matching of feature points in multimodal images," 2018, *arXiv:1810.12941*.

[31] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4353–4361.

[32] M. Chen, Y. Zhao, T. Fang, Q. Zhu, S. Yan, and F. Gao, "Geometric and non-linear radiometric distortion robust multimodal image matching via exploiting deep feature maps," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 5, no. 3, pp. 233–240, Aug. 2020.
[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.
[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: https://dblp.org/db/conf/iclr/iclr2015.html
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[39] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[41] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Neural. Inf. Process. Syst. Conf. (NIPS)*, 2017, pp. 4467–4475.
[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[43] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," 2017, *arXiv:1707.01083*.
[44] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.
[45] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[46] A. Krizhevsky. (2009). *Learning Multiple Layers of Features from Tiny Images*. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html
[47] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, and Q. V. Le, "Meta pseudo labels," 2020, *arXiv:2003.10580*.
[48] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," 2021, *arXiv:2106.04560*.
[49] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (BiT): General visual representation learning," 2019, *arXiv:1912.11370*.
[50] 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, *arXiv:2103.14030*.
[51] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 71–84. [Online]. Available: https://www.image-net.org/
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[53] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[54] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[55] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[56] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.
[57] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.
[58] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, "Thermal-visible registration of human silhouettes: A similarity measure performance evaluation," *Infr. Phys. Technol.*, vol. 64, pp. 79–86, May 2014, doi: 10.1016/j.infrared.2014.02.005.
[59] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Online mutual foreground segmentation for multispectral stereo videos," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1044–1062, Jan. 2019, doi: 10.1007/s11263-018-01141-5.
[60] J.-Y. Bouguet. *Camera Calibration ToolBox for MATLAB*. Accessed: 2004. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc
[61] S. Baker, A. Datta, and T. Kanade, "Parameterizing homographies," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RITR-06-11, 2006.
[62] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[63] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://dblp.org/db/conf/iclr/iclr2017.html
[64] D.-A. Beaupre and G.-A. Bilodeau, "Domain Siamese CNNs for sparse multispectral disparity estimation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3667–3674, doi: 10.1109/ICPR48806.2021.9412723.

INHO PARK was born in South Korea, in 1994. He received the bachelor's degree in electrical engineering from Yeungnam University, in 2019, where he is currently pursuing the degree with the Department of Electronic Engineering. His research interests include hyperspectral, hostile attack and defense, deep learning networks, and sensor fusion.

JONGMIN JEONG received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2012 and 2019, respectively. Since 2019, he has been a Senior Researcher with the Agency for Defense Development. His current research interests include computer vision, machine learning, and remote sensing image analysis.

SUNGHO KIM graduated from the College of Engineering, Korea University, in February 2000. He received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2002 and 2007, respectively. Since 2007, he has been with the Agency for Defense Development. From 2007 to 2010, he worked as a Senior Researcher with the Defense Science Research Institute (ADD). Since 2011, he has also been working as a Professor with the Department of Electronic Engineering, Yeungnam University, South Korea. His interests are hyperspectral, infrared, multi-sensor fusion, and deep learning.