IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 August 2022, accepted 27 September 2022, date of publication 3 October 2022, date of current version 10 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3211267

RESEARCH ARTICLE

Enhancing Deep-Learning Object Detection Performance Based on Fusion of Infrared and Visible Images in Advanced Driver Assistance Systems

YING-CHENG LIN, PING-YEN CHIANG, AND SHAOU-GANG MIAOU, (Senior Member, IEEE)
Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan 320314, Taiwan

Corresponding author: Shaou-Gang Miaou (miaou@cycu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2221-E-033-027.

ABSTRACT Image recognition technology plays an important role in advanced driver assistance systems (ADAS). The objective of this study is to explore the feasibility of using heterogeneous image fusion to improve the object detection performance of the ADAS. Among the many possible combinations of image types, the fusion of infrared (IR) and visible (VIS) images has great potential because of their complementary characteristics. Most studies on image fusion assume that the images involved align themselves perfectly, which is unrealistic. We address this alignment issue in this study, review various methods of image alignment and fusion, and propose an image-fusion approach that combines alignment and fusion methods for the ADAS application. Finally, we used deep learning networks to detect pedestrian and vehicle objects before and after the image fusion. The experimental results show that the fusion of IR and VIS images can improve the object detection performance of deep-learning networks. Compared with previous studies on fusion, the proposed approach ranks top if the detection accuracy improvement and execution speed are considered as a whole. This study also found that, to use image fusion to improve the object detection accuracy of deep learning networks, it is better to use fused images directly instead of unfused VIS images as the training samples.

INDEX TERMS Image fusion, infrared image, visible image, image alignment, deep learning, object detection.

I. INTRODUCTION

With the development of technology, more and more new cars are equipped with Advanced Driver Assistance Systems (ADAS). Among all relevant technology in ADAS, image recognition plays an important role because it provides image information of the surrounding environment of the vehicle for ADAS to make critical decisions that could be safety related. When the image recognition performance is excellent, it is like installing a pair of always awake and bright eyes on the vehicle to help the driver be aware of potentially dangerous situations.

The driver-assistance or self-driving vehicles encounter various situations and challenges on the road, such as safe driving and navigating in adverse weather conditions, and safe interaction with pedestrians and other vehicles [1]. To deal with such complex situations on the road, all-round sensing devices are required. For example, achieving long-range sensing on the highway is critical because it allows early detection of objects ahead and buys enough time to take proper actions like braking; in urban areas, having a wider field of view (FOV) is more important, so that pedestrians and cyclists can be detected when they are on sidewalks and crossing the road. To meet the various sensing requirements, we need multiple cameras with different FOVs. Therefore, a pair of cameras is insufficient

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

for modern ADAS or self-driving cars because most front cameras use a fixed-focus solution with a limited viewing angle and distance. In fact, the development trend of vehicle camera solutions is definitely multi-camera, and the number of cameras used can be much greater than two. For example, in 2020, Tesla had nine cameras on board: three front cameras with different FOVs, four on the sides, one on the rear, and an in-cabin camera just above the rear-view mirror. In the same year, Waymo announced that their latest vehicles had no fewer than 29 cameras on-board [2]. These cameras are installed not only at the front of the car but also at the rear and sides to provide the most complete sensing information captured from all possible views surrounding the car.

How to integrate so many cameras with different FOVs (or even modalities) and provide complementary and useful information to ADAS or self-driving cars presents many technical challenges, and one of them is the topic of interest in this study -- data fusion. The development of reliable ADAS or autonomous driving systems largely depends on the performance of automatic object detection. According to a recent report [3], intelligent multi-sensor fusion can improve the ability of self-propelled robots to detect targets. Similarly, proper multi-sensor fusion may also be a good solution to improve object detection in assisted or autonomous vehicles because it is difficult for a single sensor to generate sufficient data for accurate detection in all possible situations. Therefore, combining data from a set of heterogeneous sensors can theoretically provide richer information to produce more accurate and reliable results. However, the implementation of this theory requires experimental trials and verification.

At present, the heterogeneous sensors used in vehicles mainly include cameras, radars and LiDAR (Light Detection and Ranging). The best-known sensor for assisted or autonomous driving is the camera, which has the advantage of being cheap and reliable, but it needs to overcome poor visibility at night or in bad weather; it cannot provide range information either. In contrast, radar emits radio waves for radio detection and ranging, measuring what is reflected back by the environment. Radar also works well at night and is fairly reliable in a wide variety of weather conditions, but its sensors have a limited FOV and a much lower resolution than camera sensors. LiDAR works similarly to radar, but using invisible laser beams instead. LiDAR produces fairly detailed and extremely accurate maps of the vehicle's environment, but the technology is subject to weather conditions. Additionally, LiDAR is expensive and relatively fragile. The advantages and disadvantages of these heterogeneous sensors are summarized in Table 1. How the raw data obtained by these sensors and the information derived from them complement each other to improve driving safety is an important issue, whether it is assisted or autonomous driving. In fact, data fusion among these three types of sensors has been explored [3], [4]. In this study, we only discuss the fusion problem associated with the camera sensors capturing multiple types of images. One objective of this study is to investigate what image fusion methods are suitable to improve the recognition performance on moving objects such as pedestrians and vehicles for ADAS.

TABLE 1. The advantages and disadvantages of common heterogeneous sensors used in vehicles.

Sensor Type	Advantages	Disadvantages
Image sensor	• Cheap • High resolution • Provide complete traffic information	• Easily affected by weather, light, shadow, etc.
Millimeter-wave radar	• Not easily affected by weather • Long-range detection	• Difficult for object recognition and tracking
LiDAR	• Long measurement distance • High accuracy	• Easily affected by heavy rain, snow and dense fog • Expensive

Image fusion is an image enhancement technique that aims to combine images acquired by different types of image sensors to generate robust or informative images that can aid in subsequent processing or in decision-making. Many different types of images can be potential sources of image fusion, such as Visible (VIS), Infrared (IR), Panchromatic (PAN), Multi-Spectral (MS), Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). According to different application fields, image fusion technology can have various combination schemes, such as medical image fusion (e.g., the combination of MRI and CT) [5], telemetry image fusion (e.g., the combination of PAN and MS) [6], and heterogeneous image fusion for general applications (e.g., a combination of VIS and IR) [7].

VIS imagery in RGB has good distinguishability in human visual perception, but it is susceptible to shadows and lighting noise. In contrast, thermal IR images are less sensitive to these noises and can provide additional information for warm objects at night [8]. Therefore, fusing these two kinds of complementary images may provide a useful solution to the object detection problem. In recent years, there have been many studies on the fusion of IR and VIS images for various applications, including pedestrian detection [9], [10], [11], [12], [13], [14], face recognition [15], [16], tracking [17], [18], surveillance [19], [20], and remote sensing [21], [22]. These studies confirm that the fusion of VIS and IR images can bring some benefit in those applications, where the only one related to ADAS is the pedestrian detection, to the best of our knowledge.

However, a practical ADAS needs to detect not just pedestrian but many other objects. Therefore, we shall apply the fusion of VIS and IR images to a broader category of objects to provide a more complete study for ADAS. Specifically, the objects of interest in this study are pedestrian, automobile, bus, bicycle, and motorcycle. In addition, most of these previous studies assume that the VIS and IR images align themselves perfectly. Unfortunately, this assumption does not hold in many real-world applications. This means that there is a positional offset between the pixels corresponding to the two images. This offset problem can be due to physical characteristics of different sensors (e.g., parallax or parameter tuning), imperfect alignment algorithms, external disturbances (e.g., car shaking while driving) and hardware aging. Ignoring this real problem not only fails to achieve the expected effect of image fusion, but also may bring distortions such as ghosting artifacts on the fused image, which could be extremely detrimental to the subsequent object detection task.

In this study, we propose an image fusion approach and evaluate its benefit on the detection accuracy of moving objects commonly encountered by drivers on the road, including pedestrians and various types of vehicles. The proposed approach combines information from VIS and IR images and includes image alignment as an indispensable part of fusion for practical consideration.

The main difference between our work and some previous works are in two folds. First, we consider the object detection benefit of fusing VIS and IR images for a wider class of objects than just the pedestrian object discussed in the previous studies. Second, most existing methods operate under the assumption that the two images to be fused are perfectly aligned, and thus they can directly fuse different types of features at the corresponding pixel locations. Obviously, these methods are not only unsuitable for the unaligned images appeared in many real-world application scenarios, but also limit the further development of heterogeneous image fusion techniques for detecting moving objects on the road. This alignment issue is worthy of attention but there is still a lack of related research. Our contributions include:

- We make an important observation that image fusion in many real-world applications (including ADAS) must consider the issue of image alignment.
- We review many image fusion and image alignment techniques and propose an experimental procedure to find a good combination of these two techniques for practical applications of interest, not limit to ADAS.
- Taking the performance of object detection and the computational cost as a benchmark, we propose a relatively good approach that combines the existing image fusion and alignment methods.
- We consider the object detection benefit of fusing VIS and IR images for a wider class of objects than just the pedestrian object as discussed in the previous works, providing a more complete study.

II. RELATED WORK

We shall give a brief review on the four topics: (1) image alignment; (2) image fusion methods; (3) fusion techniques based on deep learning; and (4) road object detection based on heterogeneous images.

A. IMAGE ALIGNMENT

Image alignment technology can be divided into two broad categories: the direct (pixel-based) method and feature-based method. The former directly searches for the matching area of the pixels between the images, and then obtains the alignment relationship (through a transformation matrix) between the images from the pixels of matched area, while the latter first finds out the most discriminative feature points in the images, and then obtains the alignment relationship between the images from these points. For the direct method, if there is no preliminary alignment, the matching range of the images to be searched for a global optimal matching solution could be very large, resulting in huge computational cost. On the other hand, the feature-based method may fail when the image content has weak texture, periodicity, etc. In general, the direct method is superior to the feature-based method in terms of alignment accuracy, and it would be the opposite in terms of computational cost.

There are several well-known feature-based image alignment methods. In 2004, Lowe proposed SIFT (Scale Invariant Feature Transform) [23], which is the most famous feature detection and description algorithm in computer vision. SIFT has robust invariance to image rotation, scaling, and limited affine variations, but its main disadvantage is the high computational cost. Bay *et al.* proposed SURF (Speeded Up Robust Features) [24]. SURF features have invariance to rotation and scaling, but almost no affine invariance. Rublee *et al.* proposed ORB (Oriented FAST and Rotated BRIEF) [25], where the ORB algorithm is a mixture of the modified Features from Accelerated Segment Test (FAST) [26] and the direction standardized Binary Robust Independent Elementary Features (BRIEF) [27]. ORB features are invariant to scaling, rotation, and limited affine variations. Feature-based methods often use a combination of feature extractors (such as SIFT, SURF, and ORB) and best-feature searchers such as RANSAC (Random Sample Consensus) to align images. In general, SIFT is better than SURF, and SURF is better than ORB, in terms of the degree of extracted details of feature points. However, in terms of computing speed, it is quite the opposite.

Norm Conserved Global Affine Transformation (Norm GAT) [28] and Enhanced Correlation Coefficient (ECC) [29], [30] are two typical direct methods for image alignment. The Norm GAT algorithm has two characteristics. The first is an exact formula for maximizing the ZNCC (Zero-means Normalized Cross-Correlation). The second is to linearize the nonlinear problem, so that it does not increase the computational time complexity compared with the original GAT. ECC has been widely used in various applications, such as image registration, image mosaic, object tracking, super-resolution, visual monitoring, and medical treatment. The ECC algorithm calculates the transformation matrix by means of a motion estimation model and iteration. This method has two advantages. The first is that ECC does not produce photometric distortions that alter contrast and brightness. The second is that although the objective function is a nonlinear function of parameters (the elements of transformation matrix), the iterative scheme developed to solve the optimization problem is linear, which enables efficient implementation. In general, Norm GAT and ECC are comparable in alignment accuracy, but ECC is faster than Norm GAT in calculation speed.

Later we will show in the experiment section of this study that a simplified ECC, which is a direct method, can run even faster than feature-based methods. This surprising fact makes it the image alignment method of our choice in this study.

B. IMAGE FUSION METHODS

Image or information fusion can be divided into two broad categories: model-driven and data-driven [31]. The model-driven technique is a deterministic approach, which analyzes the target data to be fused in an analytical manner to capture the inherent features of the data, while the data-driven technique is a black-box approach, where the inherent features of the data are learned from a great quantity of relevant data, usually through a trainable neural network.

In [32], Muresan *et al.* proposed data association methods for the original data to be fused to detect and handle inconsistencies among different sensor measurements, where the sensors include a trifocal camera, a fisheye camera, a long-range radar, and LiDARs. Since human-recognizable features, such as object's appearance and motion, from various sensors are extracted, it is considered as a model-driven approach. In [33], Nie *et al.* proposed the Integrated Multimodality Fusion Deep Neural Networks (IMF-DNN) framework. The framework consists of two parts: the individual baseline neural network associated with each sensor modality and the central integrated multimodal fusion network, where the fusion can take place at all intermediate layers of each baseline neural network. How the features extracted from each modality are fused is determined by a large amount of training data, which makes IMF-DNN a data-driven fusion approach.

In our study, we exploit the characteristics of the two images obtained from two different sensors and fuse them through the process of image alignment, guided filtering, and scale decomposition, which makes it a model-driven approach. In [34], Wang *et al.* introduced "NUAN", a non-uniform attention network for multi-modal feature fusion, which is mainly a multimedia fusion technique for text, audio, and visual information. There are two main differences between the study in [34] and ours: (1) the source data to be fused are different--the former contains multimedia information, the latter contains two kinds of visual information; (2) the former is data-driven, and the latter is model-driven.

With the rapidly growing demand for various image representations, many image fusion methods for VIS and IR images have been proposed [35], where the authors provide a finer categorization than just model-driven and data-driven. A total of seven categories are considered, including multi-scale transformation, sparse representation, neural network, subspace, saliency-based methods, mixture models, and others. The main ideas of these methods are outlined below.

(1) Multi-scale transformation is a very popular technique in image fusion. In the technique, the original image is decomposed into sub-images of different scales, as different objects may show their significant characteristics at different scales. The IR and VIS image fusion scheme based on multi-scale transformation comprises three steps: performing a multi-scale transform, fused in the transform domain, and performing a corresponding inverse transform [36]. The curvelet transform [37] and the dual-tree complex wavelet transform [38] are the two representative transforms in this approach.

(2) The sparse-representation image fusion method learns a complete dictionary from a large number of high-quality natural images. An original image is then sparsely represented by this dictionary, making it possible to enhance the representation of meaningful and stable images [39]. The sparse-representation fusion method also uses a sliding window strategy to divide the original image into several overlapping patches to reduce the appearance of visual artifacts and improve the robustness of overlapping areas [40]. Yang and Li proposed Adaptive Sparse Representation (ASR) [41], which is a multi-source image fusion method based on patched signal sparse representation. To convert more interesting information from the original image into the fused image, it computes the global visual attention saliency map of the original image by analyzing all patched sparse representations. The saliency image is then used to guide the fusion rule of the local intensity of the original image.

(3) The neural network-based method imitates the perceptual behavior of the human brain to process information. Neural networks have the advantages of strong adaptability, fault tolerance, and anti-noise ability, but the neural network-based fusion method is often too computationally intensive to achieve real-time processing at present. Most neural network-based IR and VIS image fusion methods use Pulse Coupled Neural Network (PCNN) or its variants [42].

(4) Subspace-based methods, including Principal Component Analysis, Non-Negative Matrix Factorization, and Independent Component Analysis, have been successfully applied to IR and VIS image fusion [43]. Ma *et al.* proposed Gradient Transfer Fusion (GTF) [44], which is a fusion method based on gradient transfer and Total Variation (TV) minimization by mathematically transforming the fusion problem to a $\ell 1$-TV minimization problem, where the Data Fidelity Term preserves the dominant intensity distribution in the IR image, while the Regularization Term preserves the gradient variation in the VIS image.

(5) Saliency-based methods are based on the fact that it is the salient parts of the image (not individual pixels) that attract human visual attention, so saliency-based fusion methods are specifically designed to maintain the integrity of salient object region and improve the visual quality of the object in the fused image [45]. Since this is a mechanism based on the human visual system, saliency-based methods are popular for fusing IR and VIS images.

(6) The IR and VIS image fusion methods mentioned above have their own advantages and disadvantages, and a hybrid model combines their advantages to improve image fusion performance [35]. Guided Filtering Fusion (GFF) [46], [47], for example, combines multi-scale transformation with the saliency-based method.

(7) Other IR and VIS image fusion methods can inspire new ideas and directions for image fusion based on total

variation [44], fuzzy theory [48], and entropy [49], among others.

C. IMAGE FUSION BASED ON DEEP LEARNING

In recent years, deep learning has also been applied to image fusion due to the powerful image feature extraction capabilities of deep learning networks. In the past, the measurement of image activity or saliency and weight allocation in the core work of image fusion mostly relied on manual design, but a deep learning network can acquire them through training and learning to reduce the difficulty of manual design. For example, in the method proposed by Piao *et al.* [50], a Siamese network composed of Twin CNN (Convolutional Neural Network) was used. In addition, FusionGAN [51] proposed by Ma *et al.* was a method of fusing VIS and IR images through Generative Adversarial Network (GAN). The method established an adversarial game between the generator and the discriminator, in which the generator generates the fused image with IR intensity and VIS gradient, while the discriminator forces the fused image to have more VIS image details.

In [51], the authors compare their proposed method (the Siamese network) with 18 representative methods from the past. On some objective indicators of fusion performance (discussed later), their method does have a good performance. However, the performance of the CPU operation time is unsatisfactory; for example, it takes 19.47 seconds on average to fuse a pair of IR and VIS images of 270 × 360, while GFF, which is a hybrid model discussed in Section II-B, takes only 0.0899 seconds, a difference of about 217 times. Later in the experiment section, we will show that FusionGAN also has the issue of high computational complexity, while GFF ranks among the best in the combined performance of image quality metrics and execution speed.

D. HETEROGENEOUS IMAGE-BASED ROAD OBJECT DETECTION

Detecting objects on the road is an indispensable step in many driving assistance systems and has long been the focus of computer vision. Over the years, algorithms with a wide range of capabilities have been proposed, including traditional detectors [52] and the more recently dominant CNN-based detectors [11]. Recent studies have shown that heterogeneous imagery can bring great advantages, especially for computer vision covering both day and night [53]. Therefore, the release of large-scale heterogeneous image datasets [54] encouraged researchers to advance the latest technologies through the effective use of heterogeneous image data.

As a typical example of traditional approach, Hwang *et al.* proposed an extended ACF (Aggregated Channel Features) method [9], which aggregates aligned VIS and IR images for pedestrian detection through day and night. As expected, currently the more dominant approach is CNN-based [13]. For example, König *et al.* proposed an architecture that combines the Region Proposal Network (RPN) and Boosted Forest (BF) for fusing the VIS and IR information in multi-spectral images to improve automatic person detection [11]. In [12], Li *et al.* proposed an Illumination-Aware Faster Region-based CNN (IAF RCNN) and used the information from color and thermal images to improve pedestrian detection. In [55], Xu *et al.* employ a deep CNN to learn a nonlinear mapping, modeling the relations between RGB and thermal data, to improve pedestrian detection in poor lighting conditions.

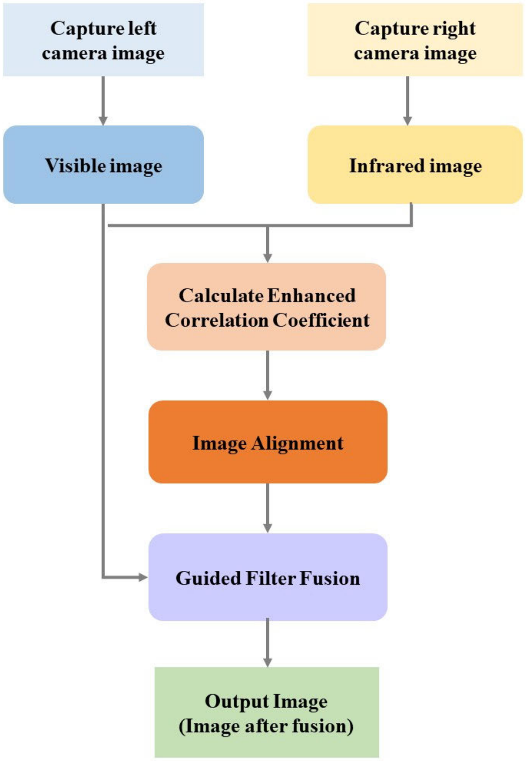

Capture left camera image
Capture right camera image
Visible image
Infrared image
Calculate Enhanced Correlation Coefficient
Image Alignment
Guided Filter Fusion
Output Image (Image after fusion)

FIGURE 1. A flowchart of the proposed fusion approach. First, a pair of visible (VIS) and infrared (IR) image is captured by an on-board dual camera, where the two corresponding lenses are located side by side. Next, image alignment is performed through the calculation of enhanced correlation coefficient. Finally, the VIS and IR images (after alignment) are fused using a guided filter, resulting in the fused output image.

III. METHOD

A. SOFTWARE AND HARDWARE EQUIPMENT AND RESOURCE FOR EXPERIMENTS

The experiment was carried out on a personal computer equipped with Windows 10 operating system, and the hardware specifications were as follows: Intel i7-8700 CPU@3.20GHz, 16G memory, NVIDIA RTX2080 GPU. For the software part, Anaconda was used with Python version 3.6.2 and PyTorch version 1.2.0. We used Python to implement image alignment and image fusion methods, and PyTorch to build a neural network for object detection.

For performance evaluation, we also used three image datasets, namely KAIST [54], TNO [56], and Pascal VOC 2012 [57]. More details on these datasets will be given in Section III-E-2.

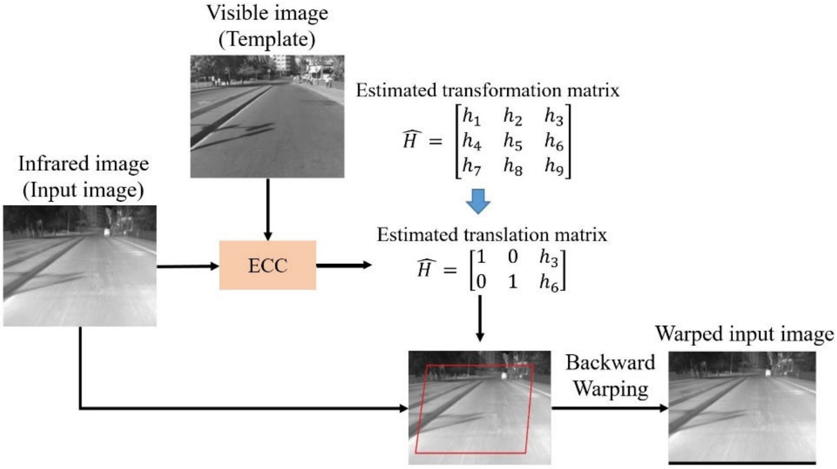

Visible image
(Template)
Estimated transformation matrix
$\widehat{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix}$
Infrared image
(Input image)
ECC
Estimated translation matrix
$\widehat{H} = \begin{bmatrix} 1 & 0 & h_3 \\ 0 & 1 & h_6 \end{bmatrix}$
Warped input image
Backward
Warping

FIGURE 2. A flowchart of our simplified ECC image alignment method (adapted from [29]). The main idea of ECC image alignment method is to estimate the transformation matrix (denoted by $\hat{H}$) of the two images involved (visible and infrared images in our case) by finding the maximum correlation coefficient between them. One image is treated as a template and the other is regarded as an input (to be transformed). The original 3 × 3 transformation matrix is simplified to 2 × 3 in this study. The IR image is aligned through the estimated transformation matrix, and the warped IR image is obtained.

B. FLOWCHART OF THE PROPOSED FUSION APPROACH

A flowchart of the proposed fusion approach is shown in Fig. 1. We used an on-board dual camera to capture both VIS and IR images simultaneously. Due to the parallax of the dual camera by itself, coupled with the car vibration and shaking during driving and other factors, the two images will have deviations in alignment. Thus, the two images are aligned first based on ECC, followed by the use of GFF.

C. IMAGE ALIGNMENT

Before fusing the VIS and IR images, image alignment is performed on these two heterogeneous images to ensure that the objects shown in the images remain at the corresponding positions in the same plane coordinates, because if the images are not aligned, it will cause artifacts in the results of subsequent image fusion. The artifacts not just affect the neural network's capability to identify objects but also produce poor human visual perception. In this study, the two cameras (VIS and IR) are put (tied together) side by side so that they usually capture almost the same scene at the same time except for a binocular disparity. So we mainly need to solve this kind of parallax problem before fusing the VIS and IR images. Since the cameras will be mounted on a vehicle moving on various road conditions in practice, a fixed calibration result may not be good enough. Therefore, we want to dynamically predict the parallax of the two cameras in the most efficient manner. A fast and accurate image alignment approach is desirable for this problem. As mentioned earlier, there are many ways to align images. Considering the computing speed, alignment accuracy and the application of this study, we believe that the ECC alignment method is a good choice for this study, as supported by the experimental results presented in Section IV.

The idea of the ECC image alignment method is to calculate the transformation matrix by finding the maximum correlation coefficient between the IR image and the VIS image, so that all pixels in the input (IR) image can be projected to the same plane coordinates as the reference (VIS) image through the transformation matrix and obtain an aligned image. Fig. 2 is a flowchart demonstrating our ECC image alignment method. In the most general ECC image alignment method, a 3 × 3 estimated transformation matrix is used to cover a variety of motion models, including Translation, Euclidean (translation + rotation), Affine (rotation + translation + scaling + shearing) and Homography (3D transformation of different planes). In the 3 × 3 estimated transformation matrix, h_1, h_5 and h_9 on the main diagonal are fixed as 1, and the values of other matrix elements depend on the motion model selected. In this study, the images are taken through two parallel and closely linked cameras and it is assumed that there are only the displacements of the X axis and Y axis, and no other changes such as rotation. So the last row of the 3 × 3 matrix (including h_7, h_8 and h_9) can be deleted to form a 2 × 3 matrix where the rotation components h_2 and h_4 are 0. The simplified matrix is helpful to improve the operation efficiency of the proposed scheme.

The ECC image alignment algorithm can be divided into the following steps:

(1) Read an IR image as the Input image and a VIS image as the Template.
(2) Convert both images to grayscale images.
(3) Select a motion estimation model.
(4) Configure the memory space to store the transformation matrix.
(5) The transformation matrix is calculated by the ECC correlation coefficient method until the set termination condition is met.

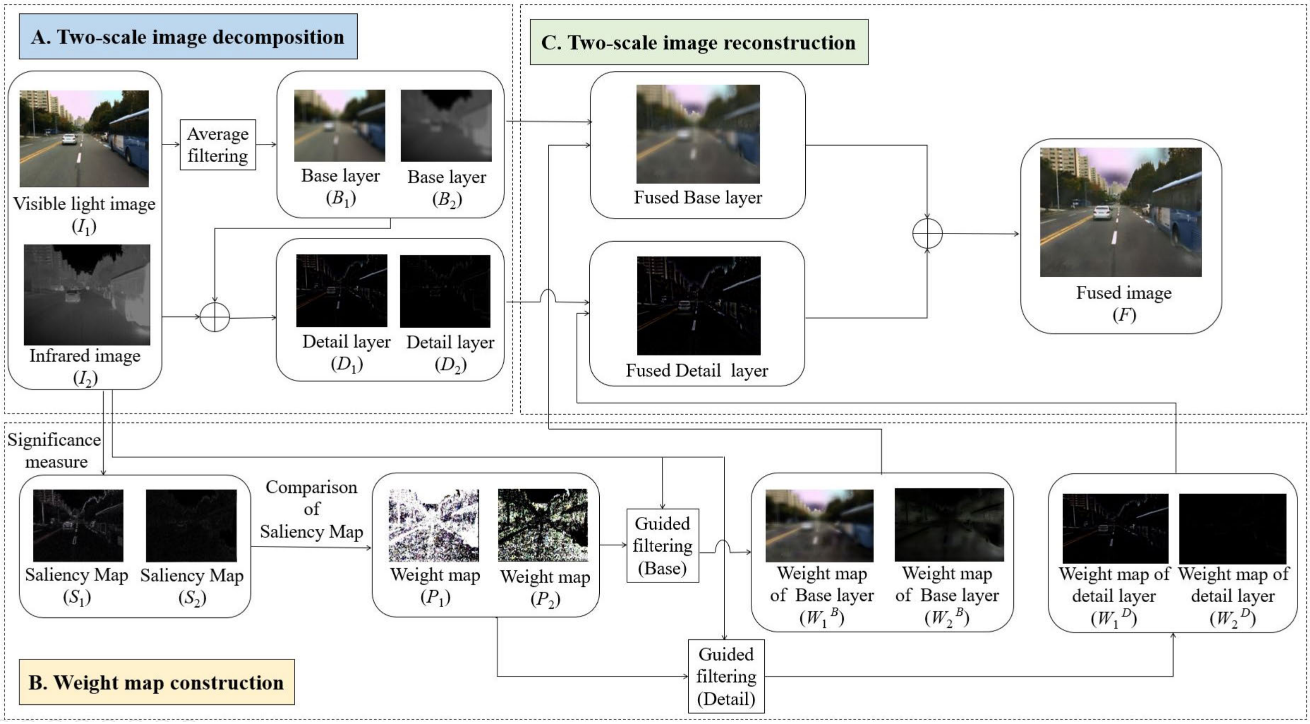


FIGURE 3. A flowchart of guided filter fusion (GFF) for street view images. It can be divided into three parts: two-scale image decomposition, weight map construction, and two-scale image reconstruction. The processing details are given in the text and the math operations involved are given from Eq. (1) to Eq. (10).

(6) The input image is aligned through the estimated transformation matrix, and the warped input image is obtained.

D. IMAGE FUSION PROCESS INVOLVING GUIDED FILITER

The guided filter has been applied to many applications, including image smoothing/enhancement, high dynamic range compression, and image defogging, etc. Here it is for image fusion. Basically, Guided Filter Fusion (GFF) [46], [47] is an image filtering technique. Through the guiding image I, the original image P is guided and filtered, so that the final output image is roughly similar to the original image P, while its texture part is similar to the guiding image I. The guided filter in GFF exploits the advantage of the bilateral filter [58] (using a non-iterative calculation method to effectively preserve the edge contour) and overcomes the shortcoming of the bilateral filter (through designing a fast filter with O(1) time complexity to ensure that gradient inversion does not occur). The guided filter not only realizes the edge smoothing function of the bilateral filter, but also performs well in the area near the detected edge. This outstanding characteristic makes GFF rank high among many representative fusion methods, as shown by the experimental result in Section IV-B. Therefore, guided filtering is our first choice for image fusion.

The flowchart in Fig. 3 shows the use of GFF. First, average filtering is performed on the two source images. The base layer and detail layer are then fused by the weighted averaging after guided filtering. The image fusion method of guided filtering can be divided into three parts, A, B and C, which are described as follows.

1) PART A: TWO-SCALE IMAGE DECOMPOSITION

The source image is decomposed into two scales by average filtering, representing the base layer of brightness and the detail layer of information, respectively. The base layer of each source image is obtained as in Eq. (1):

$$B_n = I_n * Z \tag{1}$$

where I_n is the n^{th} source image, where $n = 1, 2, \ldots, N$ (here $N = 2$ in our study, denoting the VIS and IR images), and Z is the kernel of averaging filter whose size is usually set as 31 × 31. The detail layer can be obtained by subtracting the base layer from the original image, as shown in Eq. (2):

$$D_n = I_n - B_n \tag{2}$$

The two-scale decomposition step aims at separating each source image into a base layer containing slowly-varying components and a detail layer containing higher-frequency components such as edges.

2) PART B: WEIGHT MAP CONSTRUCTION WITH GUIDED FILTERING

First, perform a saliency measure on each original image: the Laplacian filter is used to obtain the high-pass image H_n as shown in Eq. (3):

$$H_n = I_n * L \tag{3}$$

where L is a 3×3 Laplacian filter; then, the absolute value of the local mean of H_n is used to construct the saliency map S_n as given in Eq. (4):

$$S_n = \left|H_n * g_{r_g,\sigma_g}\right| \tag{4}$$

where g is a Gaussian low-pass filter of size $(2r_g+1) \times (2\sigma_g+1)$, and both parameters r_g and σ_g are usually set to 5. As stated in [45], the measured saliency maps provide good characterization of the saliency level of detail information. Then, the saliency map is examined to determine the weight map as given in Eq. (5):

$$P_n^k = \begin{cases} 1, & \text{if } S_n^k = \max\left(S_1^k, S_2^k, \ldots, S_N^k\right) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where S_n^k represents the saliency value of pixel k in the n^{th} image.

The weight map obtained above is usually noisy and does not align with the object boundary [46], [47], so if it is used directly, it may cause artifacts in the fused image. To reduce this poor impact, the idea of spatial consistency was introduced. The spatial consistency means that if two adjacent pixels have similar brightness or color, they will tend to have similar weights. A common method for spatial consistency is to formulate an energy function for the expected salient features. This energy function can encode pixel salience, modify the weight of edge alignment in a normalized manner (e.g., via smoothing), and be minimized globally to obtain the desired weight. Guided filtering uses the source image I_n as a guiding image to carry out the spatial consistency fusion for each corresponding weight map P_n. Perform guided filtering on P_n and I_n to obtain W_n^B [Eq. (6)] and W_n^D[Eq. (7)], where they are the refined weight maps generated by the base layer and the detail layer, respectively.

$$W_n^B = G_{r_1,\varepsilon_1}(P_n I_n) \tag{6}$$
$$W_n^D = G_{r_2,\varepsilon_2}(P_n I_n) \tag{7}$$

where r_1, ε_1, r_2 and ε_2 are the parameters in the process of guided filtering: r determines the salient difference of guided images under the window, and ε determines the ambiguity of guided filtering. In [59], the influences of different parameter values on the performance indexes of mean gradient, standard deviation, and mutual information of fused images are discussed. Referring to [59] and our empirical study, we set $r_1 = 30$, $\varepsilon_1 = 0.3$, $r_2 = 15$, and $\varepsilon_2 = 10^{-6}$ for the experiment section (Section IV).

3) PART C: TWO-SCALE IMAGE RECONSTRUCTION

Two-scale image reconstruction consists of the following two steps. First, the base and detail layers of source images are fused by weighted averaging, as shown in Eq. (8) and Eq. (9), respectively:

$$\bar{B} = \sum\nolimits_{n=1}^{N} W_n^B B_n \tag{8}$$
$$\bar{D} = \sum\nolimits_{n=1}^{N} W_n^D D_n \tag{9}$$

Again, here $N = 2$. Then the fused image F is obtained by combining the base layer $\bar{B}$ and the detail layer $\bar{D}$ as shown in Eq. (10):

$$F = \bar{B} + \bar{D} \tag{10}$$

E. EVALUATION OF THE BENEFITS OF IMAGE FUSION

In order to evaluate the performance of the fusion algorithm itself, we will use the evaluation indicators commonly used in image fusion. Furthermore, we will use two deep learning networks to compare the changes in the performance of object detection before and after image fusion, and explore how much the image fusion approach proposed in this study can improve the performance of object detection.

1) EVALUATION METRICS FOR IMAGE FUSION

Since it is difficult to obtain an accurate evaluation of image fusion performance only by subjective evaluation, we need fusion metrics for an objective evaluation. Many fusion metrics have been proposed, but none seems to be recognized as a de facto standard or an absolutely fair evaluation metric, so we will quantitatively evaluate the performance of fusion methods with 5 popular metrics, including entropy (EN), standard deviation (SD), structural similarity index measure (SSIM), spatial frequency (SF) and correlation coefficient (CC). A brief description of these metrics is as follows.

a. Entropy (EN)

EN measures the amount of information contained in the fused image. EN is defined in Eq. (11):

$$EN = -\sum\nolimits_{l=0}^{L-1} p_l \log_2 p_l \tag{11}$$

where L represents the total number of gray levels (set to 256 for 8-bit images) and p_l is the normalized histogram of the grayscale values in the fused image. The larger the EN, the richer the information contained in the fused image, and the better the performance of the fusion method.

b. Standard Deviation (SD)

SD is defined according to the statistical concept, which reflects the degree to which individual pixel values in an image deviate from the mean. SD is given in Eq. (12):

$$SD = \sqrt{\sum\nolimits_{i=1}^{M} \sum\nolimits_{j=1}^{N} (F\,(i,j) - \mu)^2} \tag{12}$$

where F is the fused image of size $M \times N$, and μ is the average value of the fused image F. Since areas with high contrast always attract people's attention, a fused image with high contrast usually results in a larger SD, which means that the fused image can present better visual effects.

c. Structural Similarity Index Measure (SSIM)

SSIM is mainly used to measure the similarity between the original image and its distorted image (for example, after experiencing lossy compression). In the context of image fusion, it is used to measure the structural similarity between the source image and the fused image. SSIM is mainly composed of three parts: the mean serves as an estimate of brightness, the standard deviation serves as an estimate of contrast, and the covariance is used as a measure of structural similarity. The product of these three parts is SSIM, which is expressed in Eq. (13) and Eq. (14):

$$SSIM_{X,F} = \sum\nolimits_{x,f} \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C_2} \cdot \frac{\sigma_{xf} + C_3}{\sigma_x\sigma_f + C_3} \quad (13)$$

$$SSIM = SSIM_{A,F} + SSIM_{B,F} \quad (14)$$

where $SSIM_{X,F}$ represents the structural similarity between the source image X and the fused image F; x and f represent the image blocks of the source image and the fused image in a local window of $M \times N$; σ_{xf} is the covariance of x and f; μ_x and μ_f represent the average value, σ_x and σ_f represent the standard deviation in x and f, respectively. C_1, C_2 and C_3 are the parameters that make the algorithm stable. In this study, $SSIM_{A,F}$ and $SSIM_{B,F}$ represent the structural similarity between a VIS image A and an IR image B, and their fused image F, respectively. The higher the SSIM value, the better the image fusion performance.

d. Spatial Frequency (SF)
SF is used to measure the gradient magnitude distribution of an image. The processing task to obtain SF is given as follows [from Eq. (15) to Eq. (17)]:

$$RF = \sqrt{\sum\nolimits_{i=1}^{M} \sum\nolimits_{j=1}^{N} (F(i,j) - F(i,j-1))^2} \quad (15)$$

$$CF = \sqrt{\sum\nolimits_{i=1}^{M} \sum\nolimits_{j=1}^{N} (F(i,j) - F(i-1,j))^2} \quad (16)$$

$$SF = \sqrt{RF^2 + CF^2} \quad (17)$$

where RF is the spatial row frequency, and CF is the column frequency. The larger the SF, the richer the edges and textures of the fused images.

e. Correlation Coefficient (CC)
CC is to measure the degree of linear correlation between the fused image and the source image, which is determined as follows [Eq. (18) and Eq. (19)]:

$$r_{XF} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (X(i,j) - \bar{X})(F(i,j) - \mu)}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (X(i,j) - \bar{X})^2 (\sum_{i=1}^{M} \sum_{j=1}^{N} (F(i,j) - \mu)^2)}} \quad (18)$$

$$CC = \frac{r_{AF} + r_{BF}}{2} \quad (19)$$

where X and F are the images of size $M \times N$, $\bar{X}$ is the mean of source image X, μ is the mean of the fused image F, and A and B represent the VIS image and the IR image, respectively.

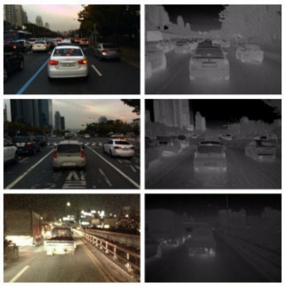


FIGURE 4. Some samples in KAIST dataset [54]: VIS images (Left) and their corresponding IR images (Right).

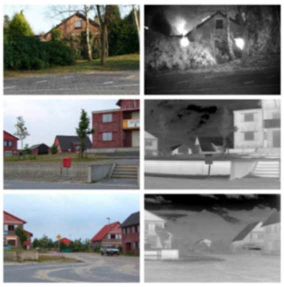


FIGURE 5. Some samples in TNO dataset [56]: VIS images (Left) and their corresponding IR images (Right).

FIGURE 6. Some samples of Pascal VOC 2012 dataset [57].

2) DATASETS USED FOR PERFORMANCE EVALUATION

For the evaluation of image fusion, this study uses the KAIST dataset [54] and the TNO dataset [56]. The former contains 95,328 pairs of VIS and IR images of size 640 × 512, and covers a variety of light conditions (including day and night cases) and weather conditions (sunny). The latter contains 34 pairs of VIS and IR images (of size 640 × 480) in the Triclobs_images folder, and all the images were taken in daytime and sunny conditions. Some samples in these two datasets are shown in Fig. 4 and Fig. 5, respectively.

In the validation phase of object detection performance, we use the Pascal VOC 2012 dataset [57], which contains approximately 15,000 labeled images classified into 20 categories. We extract a total of 7,427 images from the categories relevant to this study, including the categories of pedestrian, automobile, bus, bicycle and motorcycle. Some samples of the Pascal VOC 2012 dataset are shown in Fig. 6.

3) USE OF DEEP LEARNING NEURAL NETWORKS

After generating the fused image, we explore how much object detection performance can be improved after introducing heterogeneous image fusion. The performance evaluation will be based on commonly used and well-known indicators for general object detection, such as precision and recall. In addition, a visual (subjective) inspection is added in the experiment section, mainly by analyzing the contour of the object, exploring the possible impact on the object characteristics (especially the edge) before and after image fusion, and thus the change of detection performance. What follows is a brief review of popular object detection networks, including those used in our experiment.

Convolutional Neural Network (CNN) is the most well-known deep learning network architecture. In recent years, region-based CNN (R-CNN), Fast R-CNN, Faster R-CNN, single shot multi-box detector (SSD) and YOLO (You Only Look Once) are all derived from the CNN architecture.

Faster R-CNN [60] combines both R-CNN [61] and Region Proposal Network (RPN), which shares full-image convolution features with the detection network, greatly reducing the cost required to implement region proposals. RPN is a fully convolutional network that simultaneously predicts object boundaries and object scores at each location. RPN is trained end-to-end to generate high-quality region proposals, detected by Fast R-CNN. By sharing convolutional features, RPN and Fast R-CNN are combined into one network, and Faster R-CNN is thereby formed. However, for the embedded system, methods like Faster R-CNN require too long computing time and cannot achieve real-time performance, and this is the reason why the SSD model was proposed [62].

The SSD model eliminates bounding box and pixel or feature resampling process to achieve real-time detection speed while maintaining detection accuracy. SSD is based on a forward-propagating CNN network that produces a series of bounding boxes of fixed size, and followed by performing non-maximum suppression (NMS) to obtain the final prediction. SSD uses a single deep neural network to detect objects in images. YOLO adopts a CNN network architecture just like SSD. The characteristic of the YOLO model [63] is that it passes the image through CNN just once to get the object category and position, which greatly improves the detection speed.

Compared with YOLO, SSD has higher detection accuracy; compared with Faster R-CNN, SSD can run faster. The core feature of SSD lies in the prediction of objects and the use of scores to indicate the category to which they belong. In addition, SSD performs multi-scale predictions on several feature maps at the same time, while YOLO only performs multi-scale prediction on one feature map. The advantage of YOLO lies in a single network design, and the result of judgement will include the position of the bounding box, as well as the category and probability of each box. The entire network design is end-to-end, easy to train, and fast. In 2018, Redmon and Farhadi proposed YOLOv3 [64], which is a further optimized version of YOLOv2 (an optimized version of YOLO). After optimization, both the detection accuracy and the running speed are improved. At the same image resolution, YOLOv3 is three times faster than SSD with the same accuracy. Since SSD and YOLOv3 are two powerful and popular networks, relevant reference materials are readily available, so only a brief description about them is given above. We shall use these two networks for object detection to verify whether image fusion has resulted in the expected benefits.

IV. EXPERIMENTAL RESULTS

In this section, we perform the experiments to evaluate the proposed image fusion approach by various metrics and conduct a comprehensive comparison with other methods.

A. EVALUATION BASED ON FUSION METRICS

In this subsection, we take all 34 pairs of IR and VIS images in the TNO dataset to conduct the experiments. We evaluate the performance of existing image fusion methods, including Adaptive Sparse Representation (ASR), Curvelet Transform (CVT), Dual-tree Complex Wavelet Transform (DTCWT), Fourth Order Partial Differential Formula (FPDE), Gradient Transfer Fusion (GTF), FusionGAN, and GFF based on five evaluation metrics (EN, SD, SSIM, SF and CC). Table 2 shows the comparison results.

Table 2 shows that GFF is second to GTF and FusionGAN in the EN evaluation, second to FusionGAN in the SD evaluation, second to FusionGAN in the SSIM evaluation, second to CVT and FusionGAN in the SF evaluation, and finally second to ASR, FPDE, and FusionGAN in the CC evaluation. Although GFF ranks lower than FusionGAN in all metrics, it has the best execution time performance among all the fusion methods considered. Obviously, FusionGAN cannot meet our real-time processing requirements in terms of execution time, so we choose the second best GFF method as the image fusion method in our approach.

B. EVALUATION OF IAMGE ALIGNMENT METHODS

In this subsection, we randomly select 100 pairs of IR and VIS images of size 640 × 512 from the KAIST dataset to test currently popular image alignment methods, including SIFT+RANSAC, SURF+RANSAC, ORB+RANSAC, Norm GAT, and ECC, and the average processing time per image is given and shown in Table 3.

The results of Table 3 show that ECC is the fastest among all alignment methods. Thus, ECC is obviously a good candidate for the applications that need fast operation, such as ADAS. However, the processing speed alone cannot objectively evaluate various image alignment methods, since our fusion approach embeds the use of image alignment and we need to know further which image alignment method works best with GFF fusion, which is a special comparison benchmark of this study. For this purpose, the results of five fusion metrics (EN, SD, SSIM, SF and CC) are obtained and shown in Table 4.

TABLE 2. Performance comparison in terms of fusion metrics and execution time (the numbers associated with each method show its ranks in 6 respective performance items).

Metric / Method	EN	SD	SSIM	SF	CC	Execution time (seconds)
ASR (6, 6, 3, 5, 2, 7)	6.3196±0.5304	23.3182±1.2638	0.6126 ±0.0556	5.8737 ±0.6912	0.6462±0.0551	9.13×10^{1}
CVT (4, 4, 6, 2, 6, 4)	6.5308±0.6285	26.9374±1.3593	0.5916±0.0493	6.6722±0.6437	0.6034±0.0625	6.53×10^{-1}
DTCWT (5, 5, 5, 4, 5, 3)	6.4785±0.5945	26.2772±1.3755	0.5949±0.0577	6.5321±0.5927	0.6101±0.0564	1.30×10^{-1}
FPDE (7, 7, 4, 6, 1, 2)	6.2606±0.5453	22.8182±1.4054	0.6108±0.0571	5.6792±0.6251	**0.6513**±0.0545	9.22×10^{-2}
GTF (2, 3, 7, 7, 7, 5)	6.6877±0.5353	27.0286±1.5709	0.4741±0.0550	4.9003±0.4517	0.4854±0.0611	1.00
FusionGAN (1, 1, 1, 1, 3, 6)	**7.0618**±0.6273	**40.7618**±1.4831	**0.6349**±0.0554	**7.3528**±0.6043	0.6425±0.0548	7.16
GFF (3, 2, 2, 3, 4, 1)	6.5966±0.5134	27.2047±1.3542	0.6273±0.0465	6.6052±0.6056	0.6221±0.0499	$\mathbf{8.35 \times 10^{-2}}$

TABLE 3. Average processing time per image (640 × 512) for image alignment.

Category	Alignment Method	Execution time (seconds)
Feature -Based	SIFT+RANSAC	14.3527
	SURF+RANSAC	4.3326
	ORB+RANSAC	0.5795
Direct (pixel-based)	Norm GAT	72.6451
	ECC	**0.3415**

TABLE 4. The fusion metric performance for each image alignment method combining with GFF (the numbers associated with each method show its ranks in 5 respective performance metrics).

Index / Method	EN	SD	SSIM	SF	CC
SIFT+RANSAC (2, 3, 3, 2, 3)	6.7536 ±0.9360	27.0438 ±0.8389	0.6312 ±0.0539	6.7432 ±0.8677	0.6215 ±0.0559
SURF+RANSAC (4, 4, 4, 3, 4)	6.7342 ±0.8731	26.5834 ±0.9343	0.6257 ±0.0578	6.6324 ±0.8805	0.6148 ±0.0571
ORB+RANSAC (5, 5, 5, 5, 5)	6.7231 ±0.8618	26.2247 ±0.8615	0.6147 ±0.0572	6.6126 ±0.8433	0.5847 ±0.0555
Norm GAT (1, 1, 1, 1, 1)	**6.8253** ±0.8991	**28.6424** ±0.8851	**0.6595** ±0.0619	**7.0298** ±0.8296	**0.6423** ±0.0575
ECC (3, 2, 2, 4, 2)	6.7428 ±0.8841	27.2235 ±0.8680	0.6358 ±0.0545	6.6237 ±0.9154	0.6375 ±0.0534

The results in Table 4 show that ECC is second to Norm GAT and SIFT+RANSAC in the EN evaluation, better than ORB+RANSAC in the SF evaluation, and second to Norm GAT in the SSIM, CC and SD evaluation. On average, ECC's performance is above average, not too far behind the No. 1 Norm GAT. In contrast, the time spent by ECC is only about five thousandths of that of Norm GAT, so the overall performance of ECC is quite excellent if the results shown in both Table 3 and Table 4 are considered as a whole.

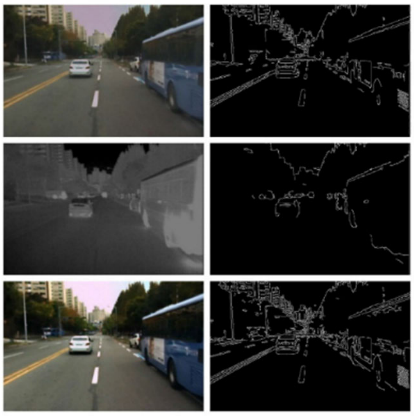


FIGURE 7. Daytime contour features. Original images (left) and their contours (right); VIS image (top); IR image (middle); Fused image (bottom).

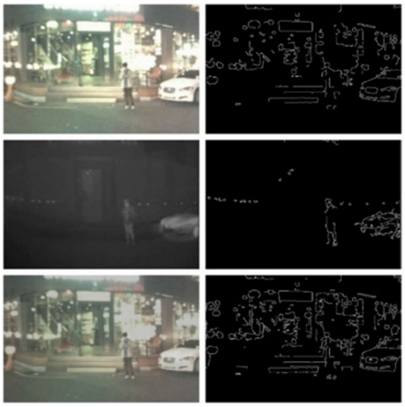


FIGURE 8. Night contour features. Original images (left) and their contours (right); VIS image (top); IR image (middle); Fused image (bottom).

C. DISCUSSION ON THE EFFECT OF IMAGE FUSION

The object detection performance of deep learning network highly depends on extracting object's features, such as point, shape, texture, color and contour. We want to actually see if fusion can bring the change of these features. Here we choose the contour feature because of its good visual effect and easy implementation, though it may not be the best one. We shall observe the changes of object contour before and after fusion. Fig. 7 and Fig. 8 show the changes of the contour features before and after image fusion during the day and at night, respectively. As shown in Fig. 7 and Fig. 8, after the

TABLE 5. Confusion matrix for alert object detection.

Actual situation Det. by network	With alert objects	Without alert objects
Objects detected	True positive (TP)	False positive (FP)
No object detected	False negative (FN)	True negative (TN)

fusion of VIS and IR images, the contours in the fused images become richer and more complete than those of separate source images, which may enable a deep learning network to provide more information for object detection, and expect to get improved performance accordingly.

To evaluate the image fusion capability of the proposed approach in object detection more accurately, we define the following terms and evaluation criteria. The case where the detection network detects an alert object within the detection range is known as positive, and the case where no alert object is detected is known as negative. The detection results can also be divided into two situations – with and without alert objects.

Depending on whether or not there really are alert objects within the detection range, we have four types of relationship as shown in Table 5. The Precision is defined as Eq. (20), which is expressed as the proportion of correctly detected alert objects to all objects detected as alert objects. The Recall is defined as Eq. (21), which is expressed as the proportion of correctly detected alert objects to all actual alert objects. The F1-score is defined as Eq. (22), which is the harmonic mean of precision and recall. Finally, the False Negative Rate (also called Miss Rate) is defined as Eq. (23), which is the proportion of false judgments out of all actual alert objects.

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$\begin{aligned} F1 - score &= \frac{2 \times Precsion \times Recall}{Precsion + Recall} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{22}$$

$$False\ Negative\ Rate(FNR) = \frac{FN}{TP + FN} \tag{23}$$

For the test images in our experiment, we provide statistics on the detection of alert objects within the threat range (the close distance in front of the vehicle and the adjacent lanes of the vehicle), as long as the alert objects are correctly detected within the range, among which the alert objects include cars, locomotives, buses, motorcycles, and pedestrians. We take 1,000 consecutive frames from each of the three Day/Night videos as experimental samples, and provide the statistics of alert object detection results according to Table 5. To evaluate the performance of our fusion approach by deep learning networks, we derive the object detection accuracy as well as the processing speed represented by the number of input frames that can be processed per second. The statistics are shown in Tables 6-9.

From the experimental results shown in Tables 6-9, it can be clearly seen that fusion processing can indeed improve the detection performance, and this improvement is achieved at a small computational cost. Figs. 9-12 show typical examples of image changes and detection results before and after fusion. These four cases show that, in general, the numbers of TP and FN cases will increase and decrease, respectively, for both object detection networks, implying that fusion has potential benefits, regardless of the object detection network used.

In addition, we can see that the performance improvement is much more significant at night than that in the daytime. Specifically, the former is about 0.22% and the latter is about 1.27% on average in terms of F1-score. This phenomenon may be due to the direct benefit from the contribution of IR images at night.

Fig.13 and Fig.14 show more cases that object detection performance is improved through fusion. Although our overall performance has improved, there are still some individual cases where some alert objects were detected successfully before fusion, but they were undetected after fusion. Examples of such cases are as shown in Fig. 15 and Fig. 16. The reason for this may be that the properties of the fused VIS image are affected by the IR image, and we used the unfused VIS images for training the object detection network, so that the confidence in recognizing the fused image decreases, resulting in very few cases where TP drops and FN increases.

To fully evaluate the strengths and weaknesses of our approach, we compare it with other published methods that also perform object detection after the fusion of IR and VIS images. The test has been conducted with six day/night videos from the KAIST dataset, and all of which were taken on urban roads. Since those previous methods consider pedestrian objects only as the basis for comparison, the following comparison is also made only for pedestrians. The comparison benchmark is the ability to reduce the false negative (missed detection) rate [Eq. (23)] through fusion before and after. The fusion methods considered include ACF+T+THOG, Halfway Fusion, Fusion RPN, Fusion RPN+BF, Illumination-aware Faster R-CNN (IAF-RCNN), Cross-Modality Interactive Attention Network (CIAN), Aligned Region CNN (AR-CNN), and the proposed approach. The corresponding object detectors associated with the above methods are shown in Table 10. The test results for the daytime situation and the night case are shown in Table 11 and Table 12, respectively. In Table 11 or 12, the relative improvement (RI) is defined as Eq. (24):

$$RI = \frac{FNR\,(before\ fusion) - FNR(after\ fusion)}{FNR(before\ fusion)} \times 100\% \tag{24}$$

TABLE 6. The detection results before and after daytime fusion in the YOLOv3 experiment.

Perf. Indexes Video clips	Precision	Recall	F1-score	Execution Speed
Day1 (before fusion)	100.0%	93.8%	96.8%	25.0 frame/s
Day1 (after fusion)	100.0%	94.2%	97.0%	23.5 frame/s
Day2 (before fusion)	100.0%	95.7%	97.8%	25.0 frame/s
Day2 (after fusion)	100.0%	96.0%	97.9%	22.8 frame/s
Day3 (before fusion)	99.4%	90.9%	95.0%	25.0 frame/s
Day3 (after fusion)	99.6%	91.4%	95.3%	21.9 frame/s

TABLE 7. The detection results before and after night fusion in the YOLOv3 experiment.

Perf. Indexes Video clips	Precision	Recall	F1-score	Execution Speed
Day1 (before fusion)	97.7%	82.0%	89.1%	25.0 frame/s
Day1 (after fusion)	98.3%	83.4%	90.2%	23.6 frame/s
Day2 (before fusion)	95.9%	75.5%	84.5%	25.0 frame/s
Day2 (after fusion)	97.5%	77.0%	86.1%	22.4 frame/s
Day3 (before fusion)	97.3%	85.9%	91.3%	25.0 frame/s
Day3 (after fusion)	98.4%	87.0%	92.3%	23.5 frame/s

TABLE 8. The detection results before and after daytime fusion in the SSD experiment.

Perf. Indexes Video clips	Precision	Recall	F1-score	Execution Speed
Day1 (before fusion)	99.9%	94.6%	97.2%	25.0 frame/s
Day1 (after fusion)	99.9%	94.8%	97.3%	22.3 frame/s
Day2 (before fusion)	99.7%	96.4%	98.0%	25.0 frame/s
Day2 (after fusion)	99.8%	96.8%	98.3%	21.7 frame/s
Day3 (before fusion)	99.6%	91.4%	95.3%	25.0 frame/s
Day3 (after fusion)	99.8%	91.8%	95.6%	21.0 frame/s

TABLE 9. The detection results before and after night fusion in the SSD experiment.

Perf. Indexes Video clips	Precision	Recall	F1-score	Execution Speed
Day1 (before fusion)	98.1%	83.0%	90.0%	25.0 frame/s
Day1 (after fusion)	98.8%	84.6%	91.2%	22.3 frame/s
Day2 (before fusion)	96.4%	77.6%	86.0%	25.0 frame/s
Day2 (after fusion)	97.2%	80.0%	87.8%	20.9 frame/s
Day3 (before fusion)	97.7%	86.7%	91.9%	25.0 frame/s
Day3 (after fusion)	98.4%	87.8%	92.8%	22.4 frame/s

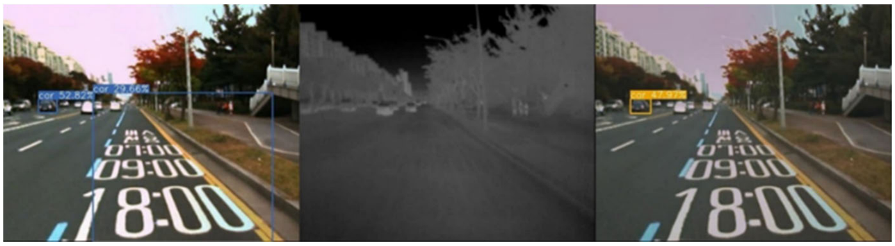


FIGURE 9. YOLOv3 object detection result. VIS image (left); IR image (middle); Fused image (right).

A zero or negative value of relative improvement indicates no improvement, whereas a higher positive value indicates better improvement.

From Table 11, we observe: (a) Halfway Fusion and AR-CNN give no improvement; (b) The image fusion approach proposed in this study gives the largest relative improvement for pedestrian object detection. From Table 12, we observe: (a) Halfway Fusion and Fusion RPN give no improvement; (b) The proposed approach is second only to CIAN in Relative Improvement metric for pedestrian

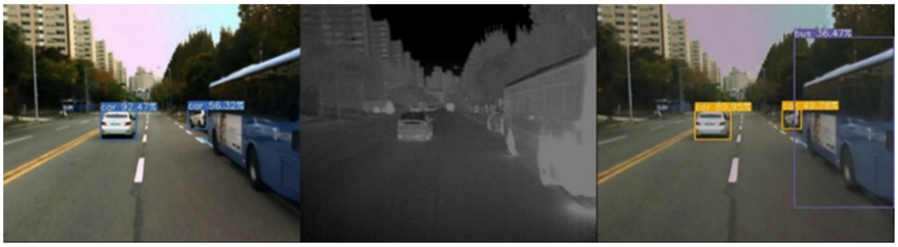


FIGURE 10. YOLOv3 object detection result. VIS image (left); IR image (middle); Fused image (right).

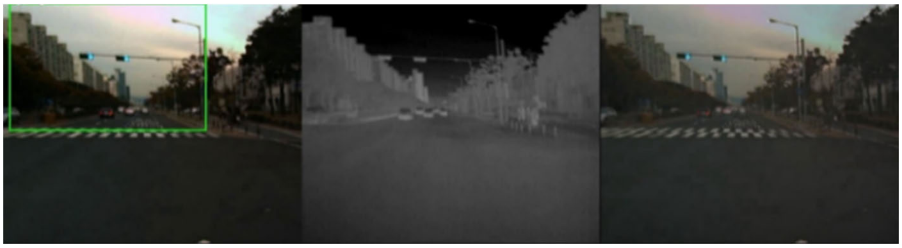


FIGURE 11. SSD object detection result. VIS image (left); IR image (middle); Fused image (right).

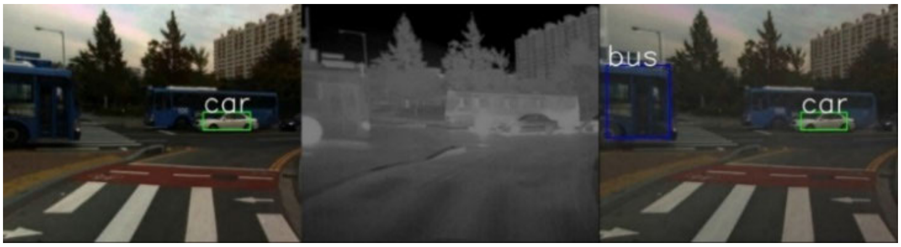


FIGURE 12. SSD object detection result. VIS image (left); IR image (middle); Fused image (right).

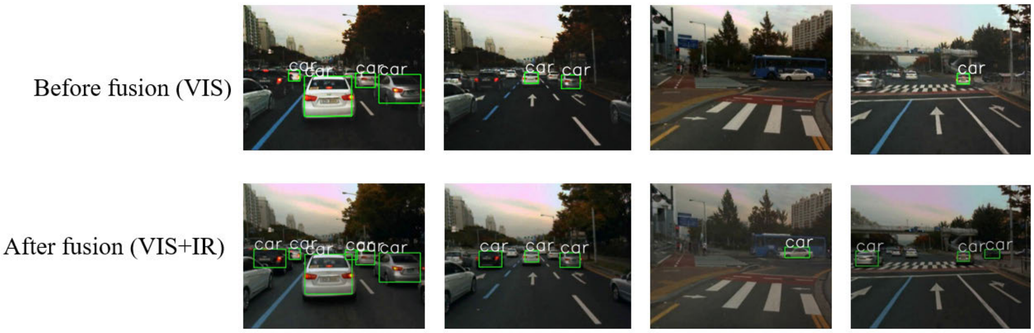


FIGURE 13. More SSD object detection results.

object detection. From Table 11 and Table 12, we conclude that CIAN and the proposed approach give the best comparable performance among others in terms of the relative improvement for pedestrian object detection. A further comparison between CIAN and the proposed approach shows that the proposed approach executes about 1.5 times faster than CIAN, as shown in Table 13.

Since there are no nighttime images in the training set and the properties of the fused image will be modified, the detection performance of the object detector trained only with daytime images before fusion may decrease in nighttime (for example, the false negative rate is high). In contrast, training with night-time fused images is expected to improve the detection performance at night. To verify this conjecture, we divide the six night videos in the KAIST

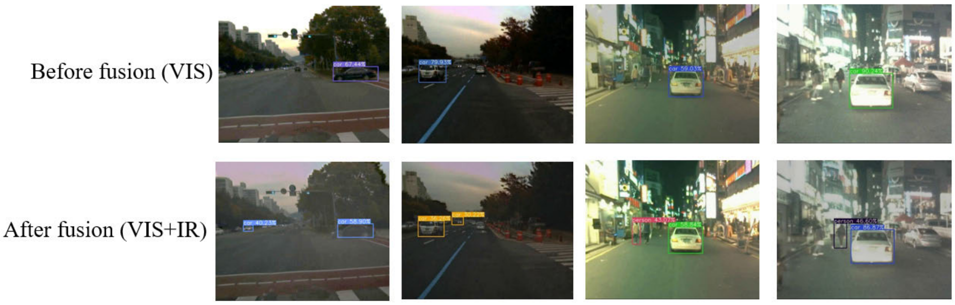


FIGURE 14. More YOLOv3 object detection results.

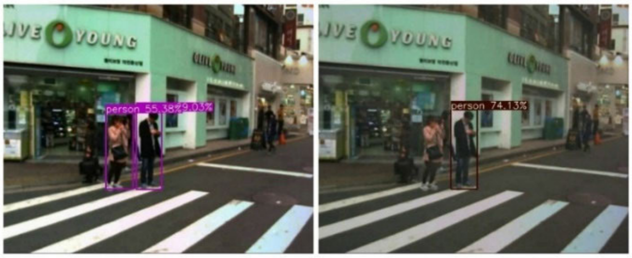


FIGURE 15. Example of performance degradation for the alert objects in daytime. (Left) image before fusion; (right) image after fusion.

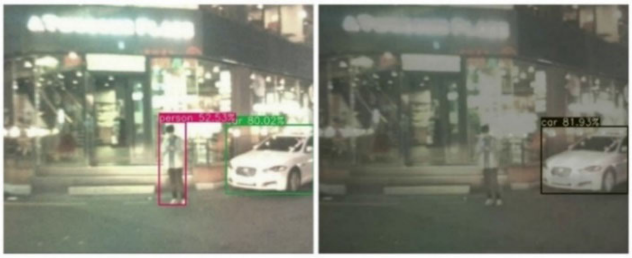


FIGURE 16. Example of performance degradation for the alert objects at night. (Left) image before fusion; (right) image after fusion.

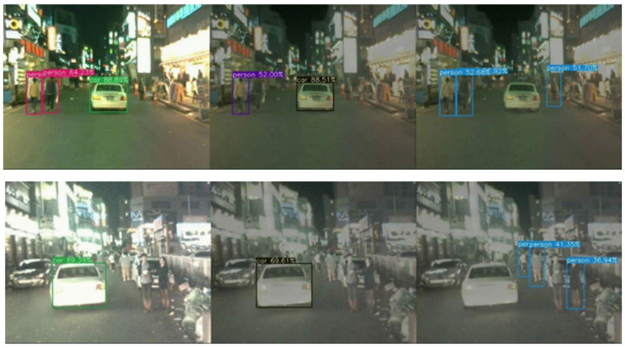


FIGURE 17. Pedestrian detection results in various sample configurations. Training set: VIS image and test set: VIS image (left); Training set: VIS image and test set: fused image (middle); Training set: fused image; test set: fused image (right).

TABLE 10. The object detectors used with fusion methods.

Fusion Method	Object Detector
ACF+T+THOG [9]	HOG (Histogram of Oriented Gradient)
Halfway Fusion [10]	Faster-RCNN
Fusion RPN [11]	RPN (Region Proposal Network)
Fusion RPN+BF [11]	RPN+BDT (Boosted Decision Trees)
IAF-RCNN [12]	Faster-RCNN
CIAN [13]	VGG-16
AR-CNN [14]	Faster-RCNN
Proposed	SSD and YOLOv3

dataset into two groups: Set 03-05 (as the training set) and Set 09-11 (as the test set), and fuse the images of these datasets through our fusion approach in advance and use YOLOv3 for training and testing. Since the KAIST dataset is mainly for pedestrian detection research, here we only report the detection results of the pedestrian object for comparison.

We provide statistics on pedestrian detection results based on the experimental samples taken from the three night videos (Set 09-11) with 1,000 consecutive frames each. The counting results of alert object (pedestrian in this case) detection by the YOLOv3 network are shown in Table 14. The precision, recall and F1-score are then derived and shown in Table 15.

It is clear from Table 15 that the overall detection performance evaluated by precision, recall, or F1-score improved after simply adjusting the training samples from ''With Visible Images Only'' to ''With Fused Images''. This proves that our conjecture is correct, that is, the properties of VIS images are indeed changed because of the contribution of IR images through image fusion. This also proves that our approach yields better results when applied to nighttime datasets when fused nighttime images are used for training. Fig. 17 shows two sets of images demonstrating the effects of using different training sample configurations.

From Fig. 17, we can observe that the network trained using the fused images can still successfully detect objects previously affected by infrared rays, and objects that were previously undetectable become detectable. Furthermore, when using image fusion to improve the object detection accuracy of a neural network, it is better to directly use the fused image rather than the unfused visible image as the training samples.

TABLE 11. Comparison of pedestrian detection performance with KAIST dataset in daytime situation.

Method	False Negative Rate (before fusion)	False Negative Rate (after fusion)	Absolute Improvement	Relative Improvement
ACF+T+THOG [9]	29.85%	29.58%	+0.27%	+0.90%
Halfway Fusion [10]	24.29%	24.88%	-0.59%	-2.43%
Fusion RPN [11]	19.69%	19.55%	+0.14%	+0.71%
Fusion RPN+BF [11]	16.60%	16.49%	+0.11%	+0.66%
IAF-RCNN [12]	14.95%	14.55%	+0.40%	+2.68%
CIAN [13]	15.13%	14.77%	+0.36%	+2.38%
AR-CNN [14]	9.91%	10.60%	-0.69%	-6.96%
Proposed	7.40%	6.94%	+0.46%	**+6.22%**

TABLE 12. Comparison of pedestrian detection performance with KAIST dataset in the night case.

Method	False Negative Rate (before fusion)	False Negative Rate (after fusion)	Absolute Improvement	Relative Improvement
ACF+T+THOG [9]	36.77%	34.98%	+1.79%	+4.87%
Halfway Fusion [10]	26.12%	26.59%	-0.47%	-1.80%
Fusion RPN [11]	21.83%	22.12%	-0.29%	-1.33%
Fusion RPN+BF [11]	15.28%	15.15%	+0.13%	+0.85%
IAF-RCNN [12]	18.26%	18.11%	+0.15%	+0.82%
CIAN [13]	12.43%	11.13%	+1.30%	**+10.46%**
AR-CNN [14]	14.21%	13.73%	+0.48%	+3.38%
Proposed	19.68%	18.68%	+1.00%	+5.08%

TABLE 13. Implementation time of CIAN and the proposed approach.

Methods	Execution time (fps)
CIAN [13]	15
Proposed	**22**

TABLE 14. The pedestrian detection results (all tested with fused images) using the YOLOv3 trained by different training samples.

Detection result Training samples	TP	FP	FN
With Visible Images Only	2291	23	516
With Fused Images	2348	12	459

TABLE 15. Performance evaluation based on the results given in Table 14.

Detection result Training samples	Precision	Recall	F1-score
With Visible Images Only	99.0%	81.6%	89.5%
With Fused Images	99.5%	83.6%	90.9%

V. CONCLUSION

In this study, we propose an approach for fusing IR and VIS images through dual cameras and explore its feasibility in ADAS by combining it with an object detection network. We used the ECC algorithm to align the IR and VIS images and used the guided filter fusion (GFF) method to fuse the two source images to obtain a fused image that retains more useful information for detection than individual source images. The proposed approach combines the details of two images through image fusion to improve the integrity of the contour of the image object, thereby improving the accuracy of object detection. It can reduce the false-negative rate in object detection and has a good execution speed compared with previous studies.

To ensure that the fusion and alignment methods we chose were a good combination, we compared our ECC alignment method with the commonly used methods (SIFT, SURF, ORB, and Norm GAT). Although the ECC method is slightly inferior to the Norm GAT method in terms of improving the fusion quality metric, it is the fastest among all the methods considered in terms of execution speed. In terms of image fusion, FusionGAN performed slightly better than the GFF fusion method used in this study for all five fusion metrics. However, it required a large amount of computing time, making it difficult, if not impossible, in real-time applications. In this regard, GFF is advantageous.

To evaluate our fusion approach, we used the YOLOv3 and SSD networks to train and detect alert objects in the images before and after image fusion, where the alert objects included cars, locomotives, buses, motorcycles, and pedestrians. The experimental results showed that the detection accuracy was improved, confirming the benefits of fusion. We also compared the results with previously published IR and VIS image fusion methods for pedestrian detection, and the results showed that the proposed approach was the best when comprehensively considering the detection accuracy and execution time performance. We also observed that the IR images changed the properties of the VIS images during fusion. Therefore, the network trained using the VIS image dataset only caused a decrease in the confidence index of object detection after fusion as well as a small decrease in TP and an increase in FN. Therefore, to use image fusion to improve the accuracy of object detection networks, it is better to use fused images directly rather than unfused visible images as the training samples.

REFERENCES

[1] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, Mar. 2021.

[2] (Apr. 6, 2020). *How Many Cameras do We Need? 29 Says Waymo, Videantis*. [Online]. Available: http://www.videantis.com/how-many-cameras-do-we-need-29-says-waymo.html

[3] S. Jusoh and S. Almajali, "A systematic review on fusion techniques and approaches used in applications," *IEEE Access*, vol. 8, pp. 14424–14439, 2020.

[4] L. Y. Hsu, "Vehicle environment sensing technology," *Automot. Res. Test.*, vol. 5, pp. 31–42, Jun. 2020.

[5] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, Sep. 2014.

[6] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, Nov. 2016.

[7] X. Zhang, P. Ye, and G. Xiao, "VIFB: A visible and infrared image fusion benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 104–105.

[8] Y. Yan, J. Ren, H. Zhao, J. Zheng, M. Z. Ezrinda, and J. Soraghan, "Fusion of thermal and visible imagery for effective detection and tracking of salient objects in videos," in *Proc. Pacific Rim Conf. Multimedia*, vol. 9917. Las Vegas, NV, USA, Jun. 2016, pp. 697–704.

[9] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[11] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 243–250.

[12] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, Jan. 2019.

[13] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, Oct. 2019.

[14] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5126–5136.

[15] H. Hariharan, A. Koschan, B. Abidi, A. Gribok, and M. Abidi, "Fusion of visible and infrared images using empirical mode decomposition to improve face recognition," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 2049–2052.

[16] S. G. Kong, I. Heo, B. R. Abidi, I. K. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition review," *Comput. Vis. Image Understand.*, vol. 97, no. 1, pp. 103–135, Jan. 2005.

[17] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Inf. Sci.*, vol. 55, no. 3, pp. 590–599, 2012.

[18] D. Smith and S. Singh, "Approaches to multi-sensor data fusion in target tracking: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 12, pp. 1696–1710, Dec. 2006.

[19] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*, vol. 4338. Madurai, India: Springer, 2006, pp. 528–539.

[20] C. O. Conaire, N. E. O'Connor, E. Cooke, and A. F. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *Proc. 9th Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–7.

[21] S. G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone, "Image fusion techniques for remote sensing applications," *Inf. Fusion*, vol. 3, no. 1, pp. 3–15, 2002.

[22] H. Li, W. Ding, X. Cao, and C. Liu, "Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing," *Remote Sens.*, vol. 9, no. 5, p. 441, 2017.

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

[24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

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

[26] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.* Graz, Austria, 2006, pp. 430–443.

[27] M. Calonder, V. Leptit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.* Heraklion, Greece, Sep. 2010, pp. 778–792.

[28] T. Wakahara and Y. Yamashita, "Acceleration of GAT correlation for distortion-tolerant image matching," in *Proc. Int. Conf. Pattern Recogn.* Tsukuba, Japan, Nov. 2012, pp. 746–749.

[29] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1858–1865, Oct. 2008.

[30] G. D. Evangelidis and E. Z. Psarakis, "Projective image alignment by using ECC maximization," in *Proc. Int. Conf. Comput. Vision Theory Appl.* Madeira, Portugal, Jan. 2008, pp. 1–8.

[31] H. Shen, M. Jiang, J. Li, C. Zhou, Q. Yuan, and L. Zhang, "Coupling Model- and data-driven methods for remote sensing image restoration and fusion: Improving physical interpretability," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 231–249, Jun. 2022.

[32] M. P. Muresan, I. Giosan, and S. Nedevschi, "Stabilization and validation of 3D object position using multimodal sensor fusion and semantic segmentation," *Sensors*, vol. 20, no. 4, p. 1110, Feb. 2020.

[33] J. Nie, J. Yan, H. Yin, L. Ren, and Q. Meng, "A multimodality fusion deep neural network and safety test strategy for intelligent vehicles," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 2, pp. 310–322, Jun. 2021.

[34] B. Wang, G. Dong, Y. Zhao, R. Li, Q. Cao, and Y. Chao, "Non-uniform attention network for multi-modal sentiment analysis," in *Proc. Int. Conf. Multimedia Modeling*, Jun. 2022, pp. 612–623.

[35] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[36] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fusion*, vol. 4, no. 4, pp. 259–280, 2003.

[37] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[38] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[39] Q. Zhang, Y. Liu, R. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, Apr. 2018.

[40] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[41] B. Yang and S. Li, "Visual attention guided image fusion with sparse representation," *Optik*, vol. 125, no. 17, pp. 4881–4888, Sep. 2014.

[42] T. Xiang, L. Yan, and R. Gao, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking PCNN in NSCT domain," *Infr. Phys. Technol.*, vol. 69, pp. 53–61, Mar. 2015.

[43] D. P. Bavirisetti, G. Xiao, and G. Liu, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–9.

[44] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[45] X. Zhang, Y. Ma, F. Fan, Y. Zhang, and J. Huang, "Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 8, pp. 1400–1410, 2017.

[46] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[47] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[48] S. Rajkumar and P. C. Mouli, "Infrared and visible image fusion using entropy and neuro-fuzzy concepts," in *Proc. ICT Crit. Infrastruct., 48th Annu. Conv. Comput. Soc. India*, 2014, pp. 93–100.

[49] J. Zhao, G. Cui, X. Gong, Y. Zang, S. Tao, and D. Wang, "Fusion of visible and infrared images using global entropy and gradient constrained regularization," *Infr. Phys. Technol.*, vol. 81, pp. 201–209, Mar. 2017.

[50] J. Piao, Y. Chen, and H. Shin, "A new deep learning based multi-spectral image fusion method," *Entropy*, vol. 21, no. 6, pp. 1–16, Jun. 2019.

[51] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[52] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1751–1760.

[53] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6983–6991.

[54] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.

[55] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5363–5371.

[56] A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, Dec. 2017.

[57] (2012). *Visual Object Classes Challenge 2012 (VOC2012)*. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

[58] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, Jul. 2002.

[59] Z. Wang and Y. Zhu, "Multi-focus image fusion based on the improved PCNN and guided filter," *Neural Process. Lett.*, vol. 45, no. 1, pp. 75–94, 2017.

[60] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

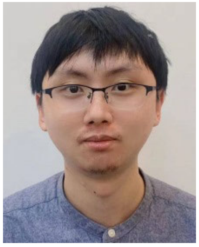
[61] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2014, pp. 580–587.

[62] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, Oct. 2016, pp. 21–37.

[63] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[64] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

YING-CHENG LIN was born in Nantou City, Taiwan, in 1991. He received the B.S. and M.S. degrees from the Department of Electronic Engineering, Chung Yuan Christian University (CYCU), Taiwan, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electronic engineering. His current research interests include image processing and deep learning.



PING-YEN CHIANG received the B.S. degree from the Department of Computer Science and Information Engineering, Chinese Culture University, Taiwan, in 2018, and the M.S. degree from the Department of Electronic Engineering, Chung Yuan Christian University, Taiwan, in 2020. His current research interests include image processing and deep learning.

SHAOU-GANG MIAOU (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Florida, USA, in 1993. He is currently a Professor with the Department of Electronic Engineering, Chung Yuan Christian University (CYCU), Taiwan, where he was worked as the Dean of the College of Electrical Engineering and Computer Science, from 2017 to 2022. His research interests include image processing, pattern recognition, deep learning, and digital signal processing.