Deep Learning-Enhanced Multi-Modal Sensing Platform for Robust Human Object Detection and Tracking in Challenging Environments

Peng Cheng , Zinan Xiong , Yajie Bao , Ping Zhuang , Yunqi Zhang , Erik Blasch , Genshe Chen *

Posted Date: 11 July 2023

doi: 10.20944/preprints202307.0664.v1

Keywords: Human Object Recognition and tracking; Multi-Modal Sensing; EO/IR; Radar; Mobile Platform; Deep Learning; Image Fusion, Autonomous Vehicles

Article

Deep Learning-Enhanced Multi-Modal Sensing Platform for Robust Human Object Detection and Tracking in Challenging Environments

Peng Cheng [1], Zinan Xiong [1], Yajie Bao [1], Ping Zhuang [1], Yunqi Zhang [1], Erik Blasch [2] and GensheChen [1,*]

[1] Intelligent Fusion Technology, Inc., Germantown, Maryland, USA; peng.cheng@intfusiontech.com
[2] MOVEJ Analytics, Dayton, OH, erik.blasch@gmail.com
* Correspondence: gchen@intfusiontech.com

Abstract: In modern security situations, tracking multiple human objects in real-time within challenging urban environments is a critical capability for enhancing situational awareness, minimizing response time, and increasing overall operational effectiveness. Tracking multiple entities enables informed decision-making, risk mitigation, and the safeguarding of civil-military operations to ensure safety and mission success. This paper presents a multi-modal electro-optical/infrared (EO/IR) and radio frequency (RF) fused sensing (MEIRFS) platform for real-time human object detection, recognition, classification, and tracking in challenging environments. By utilizing different sensors in a complementary manner, the robustness of the sensing system is enhanced, enabling reliable detection and recognition results across various situations. Specifically designed Radar tag and thermal tag can be used to discriminate friendly and non-friendly objects. The system incorporates deep learning-based image fusion and human object recognition and tracking (HORT) algorithms to ensure accurate situation assessment. After integrating into an all-terrain robot, multiple ground tests were conducted to verify the consistent HORT in various environments. The MEIRFS sensor system has been designed to meet the Size, Weight, Power, and Cost (SWaP-C) requirements for installation on autonomous ground and aerial vehicles.

Keywords: human object recognition and tracking; multi-modal sensing; EO/IR; radar; mobile platform; deep learning; image fusion; autonomous vehicles

1. Introduction

Autonomous vehicles, including unmanned aerial vehicles (UAVs) [1–3] and unmanned ground vehicles (UGVs) [4], have found extensive applications in search and rescue due to their mobility and operational simplicity. One significant capability desired in these search and surveillance scenarios is the ability of autonomous vehicles to recognize human subjects' actions and respond accordingly. Electro-optical (EO) cameras have become essential tools on UAV and UGV platforms to enhance situational awareness, perform object detection, and enable efficient tracking capabilities. Cameras provide valuable visual information that aid in various applications, including search and rescue operations, surveillance missions, and security monitoring.

However, recognizing human objects from videos captured by a mobile platform presents several challenges. The articulated structure and range of possible poses of the human body make human object recognition and tracking (HORT) a complex task. Humans exhibit diverse movements and postures, making it difficult for an autonomous system to accurately recognize and track them in video footage. Additionally, the quality of the captured videos further complicates the recognition and classification process. Videos may suffer from perspective distortion, occlusion (when parts of the human body are obstructed), motion blur, or poor visibility in adverse weather conditions like fog or rain.

Addressing these video exploitation challenges requires advanced computer vision and image processing techniques. Researchers and engineers employ various approaches, including deep learning (DL)-based methods [5], to develop algorithms capable of robust HORT. These HORT algorithms leverage DL models trained on extensive datasets to recognize and track human subjects despite the mentioned challenges. They take into account pose variations, occlusion handling, motion estimation, and visibility enhancement techniques to improve accuracy and reliability. To quickly respond to variations, deployed systems seek sensor fusion on edge platforms [6] or with communication support from fog services [7].

Although DL-based human object detection methods have shown great potential in improving detection accuracy with high-quality images, they may encounter difficulties in challenging environments where image quality is significantly degraded. These environments include scenarios such as completely dark tunnels or situations with limited visibility due to heavy fog. In such challenging conditions, the performance of ML algorithms can be hindered. The lack of sufficient illumination in dark environments can lead to poor image quality, making it challenging for the algorithms to extract relevant features and accurately detect human objects. Similarly, heavy fog or other atmospheric conditions can cause image distortion, reduced contrast, and blurred edges, further impeding the performance of the detection algorithms.

To address these video exploitation environmental limitations, researchers are exploring alternative sensing technologies and approaches that can complement or enhance DL-based methods [8]. For example, thermal imaging sensors, such as infrared (IR) cameras, can operate effectively in low-light or no-light environments by detecting the heat signatures emitted by objects, including humans [9]. IR sensing allows for improved object detection even in complete darkness or challenging weather conditions. Radar [10–12] and Lidar technologies [13–15] also play a significant role in various tasks related to object detection. Particularly, the emergence of radar sensors holds great promise for real-time human object detection applications from moving vehicles [16,17], multimodal systems [18,19], and passive sensing [20,21]. In comparison to visible cameras, Radar and Lidar sensors may have limitations in providing detailed texture information of the viewing scene. Consequently, the lack of texture information makes it challenging to utilize radar and lidar effectively for tasks such as human object detection and recognition.

In addition to the challenges posed by HORT faces an even greater challenge when it comes to identifying (e.g., friendly/non-friendly) and continuously tracking human subjects in videos captured from platforms such as UAVs. While human detection focuses on locating individuals in a single frame, tracking requires maintaining their classification and trajectory across multiple frames. Tracking algorithms must address the inherent difficulties in accurate and reliable human detection, as errors or inaccuracies in the initial detection phase can propagate and adversely affect the tracking process. Tracking algorithms also need to handle situations where detection may fail or produce false positives, leading to incorrect associations or track drift.

This paper presents the development of a *multi-modal EO/IR and RF fused sensing* (MEIRFS) platform for real-time human object detection, recognition, and tracking in challenging environments. By utilizing different sensors in a complementary manner, the robustness of the sensing system is enhanced, enabling reliable detection results across various situations. The system incorporates DL-based image fusion and HORT algorithms to ensure accurate detection and tracking of human objects. The sensor system consists of two subsystems: 1) the EO/IR subsystem, which detects and locates human objects in line-of-sight (LOS) scenarios and continuously tracks the selected objects; 2) the RF subsystem, featuring a linear frequency modulated continuous wave (LFMCW) ranging radar and smart antenna, designed to detect and locate both LOS and non-line-of-sight (NLOS) friendly objects. These two subsystems have been successfully integrated, establishing communication between the sensor system and the host computer, thereby enabling real-time HORT of friendly human objects.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

2. System architecture

Figure 1 illustrates the complete structure of the MEIRFS sensor system designed for human object detection, recognition, and tracking. The edge platform (UAV or UGV) is equipped with all the required sensors for detecting and continuously tracking human objects. These sensors include the ranging radar, EO/IR camera, laser range finder, differential barometer, and a pan/tilt platform. Additionally, the friendly object is equipped with an IR emitter and an RF transponder, enabling easy recognition by the MEIRFS system amidst all the detected human objects.

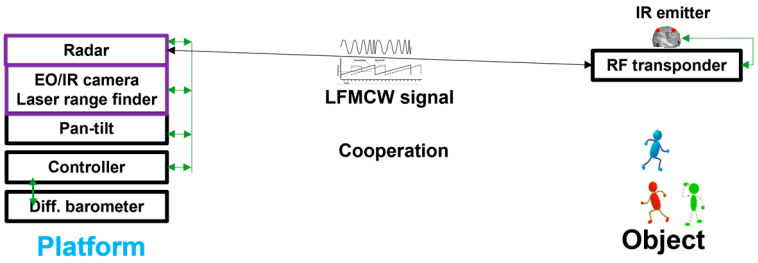

Figure 1. Structure of multimodal EO/IR and RF based sensor system.

2.1. Radio Frequency (RF) subsystem

The RF subsystem is comprised of a LFMCW ranging radar, with the transceiver located on the platform and the transponder situated on the friendly object. Additionally, a smart antenna is positioned on the platform side. The LFMCW transceiver, illustrated in Figure 2a, consists of a LFMCW transmitter, a LFMCW receiver with frequency/range scanning capability, and a signal processor. The RF system incorporates a smart antenna capable of estimating the angle between the platform and the friendly object. The smart antenna achieves a measurement accuracy of 0.8° and effectively suppresses multipath signals reflected from the ground, walls, and ceilings. Figure 2b displays the radar transponder situated on the friendly object side. The entire radar subsystem underwent testing in an indoor environment, as depicted in Figure 2c that showcases the measured distance between the platform and the friendly object. The results demonstrate the consistent detection and accurate distance measurement capabilities of the MEIRFS self-developed radar subsystem.

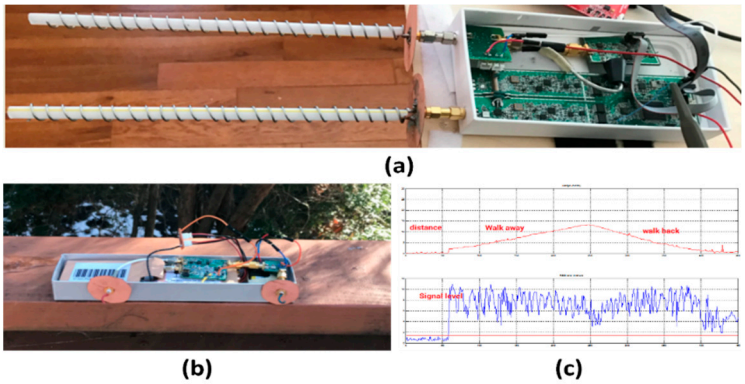

Figure 2. LFMCW Radar with smart antenna. (a) Radar transceiver on the platform side, (b) Radar transponder on the friendly object side, and (c) distance measurement in experiment.

To enhance the signal-to-noise ratio (SNR) and range detection, several techniques were implemented:

1. The RF signals were sampled multiple times, typically eight samples, and Fast Fourier Transform (FFT) calculations were performed on each sample. The results were then averaged, improving the SNR and extending the detection range.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

2. Due to varying hardware gain responses across the baseband spectrum, it was necessary to determine the local signal noise floor as a reference. By comparing the real signal with the local noise floor instead of the entire baseband noise floor, accurate detection can be achieved.
3. Local averaging windows were utilized to establish the appropriate reference level, contributing to improved detection accuracy.

The current radar range cutoff stands at just over 27 meters. If required, parameters can be adjusted to enable a longer detection range. The distance measurement update rate is set at 7 times per second. At this refresh rate, the average current draw is 700mA at 6V. The refresh rate can be increased if certain radar functions are not turned off between each update to conserve power.

The capabilities of the MEIRFS RF subsystem were tested and verified in both outdoor open environments and wooded areas. Furthermore, it was confirmed that the RF subsystem consistently detects the distance of human objects equipped with radar transponders, even through multiple drywalls.

2.2. EO/IR subsystem

The EO/IR subsystem comprises an EO camera, an IR camera, a laser rangefinder situated on the platform side, a controllable IR emitter on the friendly object side, and a pan/tilt platform. Within the subsystem, the EO camera utilizes a 3D stereo camera for visible image acquisition and depth sensing, while the long-wavelength IR camera is employed for thermal detection. Two options of IR cameras are available, allowing for interchangeability to accommodate different detection ranges. Both options have undergone comprehensive testing and successful implementation.

Aligned with the viewing direction of the IR camera, the laser rangefinder is capable of measuring distances up to 100 meters. The IR subsystem consistently distinguishes between LOS friendly and non-friendly objects by analyzing the IR signal emitted from the IR emitter equipped by the friendly object.

The hardware arrangement of the IR subsystem is depicted in Figure 3a. Both the IR camera and the laser rangefinder are aligned to point in the same direction and are mounted on the pan/tilt platform, allowing for rotation in various directions. The laser rangefinder is utilized to measure the distance of the object located at the center of the IR image's field of view. As shown in Figure 4b, the process begins with the capture of the first image t_1 from the IR camera, which detects the human object. The object's position within the IR image's field of view is then calculated. Subsequently, the lateral angle position α and the vertical angle position ϕ of the object relative to the IR camera's pointing direction can be determined. These calculated angle positions are then sent to the pan/tilt platform, which adjusts the IR subsystem's orientation to center the object within the IR camera's field of view. Thus, at time instant t_2, the distance of the object can be measured using the laser rangefinder. Figure 4c presents the flowchart illustrating the working principle of the EO/IR subsystem, highlighting its functionality in detecting, tracking, and measuring the distance of the object of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

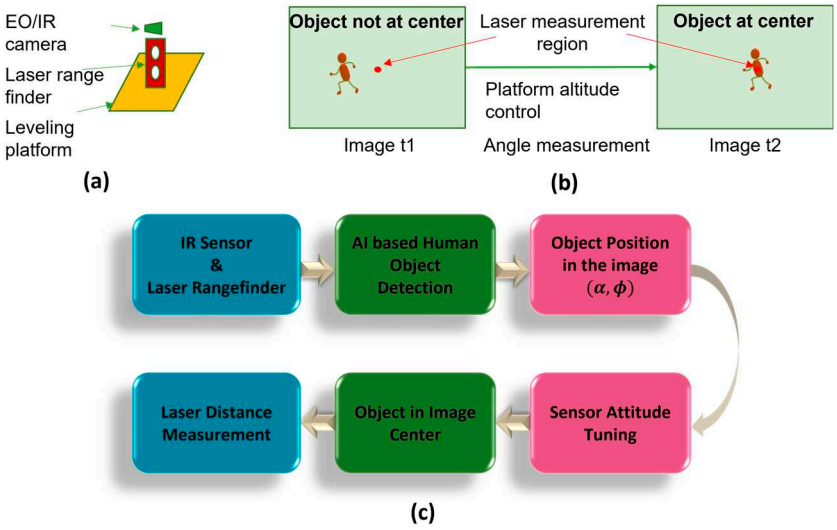

Figure 3. IR subsystem for human object detection and tracking. (a) IR subsystem hardware setup; (b) centering on interested object for distance measurement, and (c) flowchart of the working principle for the IR subsystem.

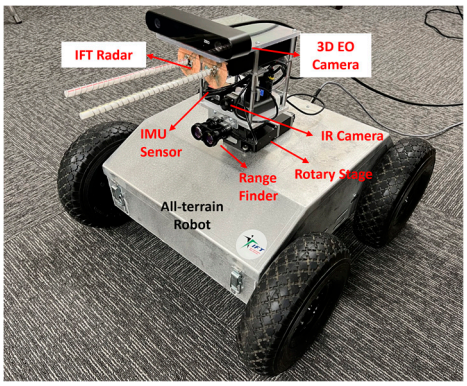

Figure 4. Assembled multimodal EO/IR and RF fused sensing system.

2.2.1. Electro-Optical (EO) camera

The 3D stereo camera from Stereolabs is used as the EO camera for both visible image acquisition and depth sensing. The camera offers advanced depth sensing capabilities and is widely used for applications such as robotics, virtual reality, autonomous navigation, and 3D mapping. Some key features of this 3D camera include: high resolution (1920×1080 pixels) visible image, depth sensing, real-time 3D mapping, and comprehensive software development kit (SDK).

In our specific application, we utilize the image captured by the left camera of the 3D stereo camera as the EO image. The left image serves as the basis for human object detection and tracking using visible light. By leveraging the visible light spectrum, one can benefit from the detailed texture information and visual cues present in the EO image, enabling accurate detection and tracking of human subjects.

2.2.2. Infrared (IR) camera

The IR subsystem incorporates two different IR cameras for varying human object detection ranges: the 9640P IR camera from ICI and the Boson 320 IR camera from Teledyne. The selection and testing of these cameras were performed to adapt to different detection requirements.

The short-range Boson 320 IR camera boasts a compact size of 21×21×11 mm and weighs only 7.5g. It is equipped with a 6.3mm lens and offers a horizontal field of view (FOV) of 34°. This camera is capable of detecting human objects up to a range of 25 meters. It features exceptional thermal

sensitivity, equal to or less than (≤) 20 mK, and an upgraded automatic gain control (AGC) filter that enhances scene contrast and sharpness in all environments. With a fast frame rate of up to 60Hz, it enables real-time human object detection. The image resolution of this camera is 320×256 pixels, and the image stream is transferred in real-time from the camera to the host PC via a universal serial bus (USB) cable.

On the other hand, the long-range ICI 9640p is a high-quality thermal grade IR camera with an image resolution of 640×512 pixels. It utilizes a 50mm athermalized Lens, providing a FOV of 12.4° × 9.3°, and has a total weight of 230g. This ICI IR camera achieves a detection range exceeding 100 meters. The maximum frame rate supported by this camera is 30Hz.

By incorporating both the Boson 320 and the ICI 9640p cameras into the IR subsystem, the MEIRFS system can adjust to different detection ranges, ensuring flexibility and adaptability in various scenarios.

2.2.3. Laser rangefinder

To overcome the limitation of the IR camera in measuring the distance of detected objects, we integrated a laser rangefinder SF30/C from Lightware into our system. The laser rangefinder is specifically designed to provide accurate distance measurements. It is aligned with the viewing direction of the IR camera, and both devices are mounted on a rotary stage. The collocated configuration ensures that the laser rangefinder is always directed towards the center of the IR camera's field of view.

When a human object of interest is detected in the FOV, the rotary stage automatically adjusts the orientation of the IR subsystem to the center of the object; affording the precise position of the object relative to the platform of the sensor system. By combining the information from the IR camera, which provides the location of the object, and the laser rangefinder, which provides the distance measurement, MEIRFS can accurately determine the spatial coordinates of the human object in real-time.

2.3. Sensor system integration

The proposed MEIRFS system is designed to be versatile and applicable to both UAVs and UGVs for various tasks. In this paper, we demonstrate the successful integration and mounting of the MEIRFS system onto an all-terrain robot platform to conduct ground tests.

By deploying the MEIRFS system on a UGV, the performance and capabilities are evaluated in real-world scenarios encountered by ground-based robotic platforms. The all-terrain robot platform provides a suitable environment for testing the MEIRFS system's functionalities such as human object detection, recognition, and tracking. These tests help validate the effectiveness and robustness of the MIERFS system in different operational conditions.

The MEIRFS integration onto the all-terrain robot platform enables us to assess the MEIRFS system's performance in practical ground-based applications, paving the way for potential deployment on both UAVs and UGVs for diverse tasks such as surveillance, search and rescue, and security operations.

2.3.1. Hardware system assembly

As shown in Figure 4, the MEIRFS system includes the following components:

1) The Radar sensor developed by Intelligent Fusion Technology, Inc.
2) The 3D EO camera from Stereolabs.
3) The FLIR Boson 320 IR camera from Teledyne.
4) The SF30/C laser rangefinder from Lightware.
5) The HWT905 inertial measurement unit (IMU) sensor from Wit-Motion.
6) X-RSW series motorized rotary stage from Zaber.
7) The IG42 all-terrain robot from SuperDroid Robots.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

To ensure an organized and compact design, all the cables of the MEIRFS system are carefully managed and extended to the interior of the robot. Inside the robot, two 12V batteries are utilized to generate a 24V DC power supply, which is required for operating both the rotary stage and the robot's wheels.

In terms of connectivity, a single USB cable is all that is necessary to establish communication between the MEIRFS system and the host computer. The USB cable connects to a USB hub integrated into the robot, facilitating seamless communication between the host computer and all the sensors, as well as the rotary stage. By consolidating the cables and employing a simplified connection scheme, the MEIRFS system ensures efficient and streamlined communication, minimizing clutter and simplifying the setup process. The organized arrangement enhances the overall functionality and practicality of the system during operation.

2.3.2. Software package

To facilitate user control and provide a comprehensive display of the detection results, a graphical user interface (GUI) software package was developed. The MEIRFS GUI software serves as a centralized platform for communication and control between the host computer and all the hardware devices in the sensor system.

The GUI software, illustrated in Figure 5, enables seamless communication and data exchange with the various components of the sensor system. The GUI acts as a user-friendly interface for controlling and configuring the system, as well as displaying key data and detection results in a clear and organized manner. Through the GUI software, users can conveniently interact with the sensor system, adjusting settings, initiating detection processes, and monitoring real-time data. The software provides an intuitive and efficient means of accessing and managing the functionalities of the MIERFS system. Specifically, the GUI software has been developed with the following capabilities:

1) Display the image acquired from EO/IR cameras.
2) Configure the machine learning model for human object detection.
3) Receive and display the measurement results from the IMU sensor.
4) Receive and display the measurement results from the Laser rangefinder.
5) Send the control command to the rotary stage.
6) Receive and display the measurement results from the radar.
7) Perform real-time object tracking to follow the interested object.

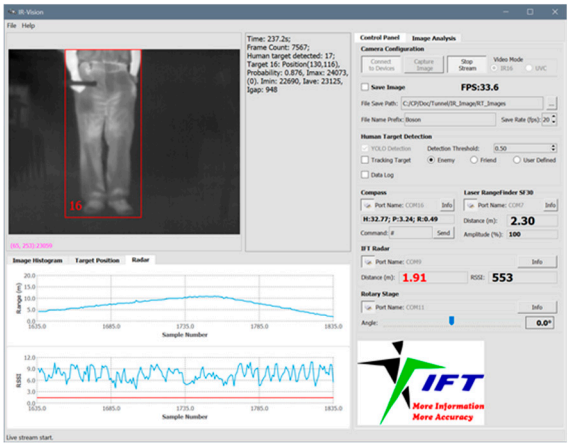

Figure 5. Assembled multimodal EO/IR and RF fused sensing system.

The measurement results from the various sensors in the MIERFS system are transmitted to the host computer at different data update rates. To ensure accurate tracking of the object, these measurements are synchronized within the GUI software to calculate the object's position. In the MIERFS system, the IR camera plays a crucial role in human object detection, recognition, and tracking. Therefore, the measurements from other sensors are synchronized with the update rate of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

the IR camera. During our testing, the real-time human object detection process achieves a continuous frame rate of approximately 35 frames per second (fps) when the laptop computer (equipped with an Intel Core i9-11900H CPU and Nvidia RTX-3060 laptop GPU) was connected to a power source. When the laptop computer operated solely on battery, the frame rate reduced to about 28 fps. Each time a new frame of the IR image is received in the image acquisition thread, the software updates the measured data from all the sensors. The synchronization ensures that the measurement results from different sensors are aligned with the latest IR image frame, providing accurate and up-to-date information for human object detection and tracking.

3. Enhance the MIERFS system with deep learning methods

3.1. Deep Learning-based algorithm for human object detection

After evaluating various DL-based object detection algorithms suitable for real-time applications [22,23], we selected the open-source YOLOv4 (You Only Look Once) detector [5] as the tool for EO/IR image analysis in human object detection. The YOLOv4 detector is recognized as one of the most advanced DL algorithms for real-time object detection. It employs a single neural network to process the entire image, dividing it into regions and predicting bounding boxes and probabilities for each region. These bounding boxes are weighted based on the predicted probabilities.

The YOLOv4 model offers several advantages over classifier-based systems. It considers the entire image during testing, leveraging global context to enhance its predictions. Unlike systems such as the region-based convolutional neural network (R-CNN), which require thousands of network evaluations for a single image, YOLOv4 makes predictions in a single evaluation, making it remarkably fast. In fact, it is over 1000 times faster than R-CNN and 100 times faster than Fast R-CNN [5].

To ensure the YOLOv4 detector's effectiveness in different scenarios, we gathered more than 1000 IR images, encompassing various cases, as depicted in Figure 6. Additionally, we considered scenarios where only a portion of the human body was within the IR camera's field of view, such as the lower body, upper body, right body, and left body. Once the raw IR image data was annotated, both the annotated IR images and their corresponding annotation files were used as input for training the YOLOv4 model. The pre-trained YOLOv4 model, initially trained with the Microsoft common objects in context (COCO) dataset, served as the starting point for training with the annotated IR images.

Figure 6. Examples of training image data obtained in different environments.

Once the training of the YOLOv4 model was finalized, we proceeded to evaluate its performance using IR images that were not included in the training process. Figure 7 showcases the effectiveness of the trained YOLOv4 model in accurately detecting human objects across various scenarios, including:

1) Human object detection in indoor environments.
2) Human object detection in outdoor environments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

3) Detection of multiple human objects within the same IR image.
4) Human object detection at different distances.
5) Human object detection regardless of different human body gestures.

The trained YOLOv4 model exhibited satisfactory performance in all of these scenarios, demonstrating its ability to robustly detect human objects in diverse environments and under various conditions.

Figure 7. Human object detection results with trained YOLOv4 model.

3.2. Sensor Fusion and Multi-target Tracking

Although the IR image alone is effective for human object detection, it may not provide optimal performance in multiple human objects tracking tasks due to its limited color and texture information compared to visible light images. To address this limitation and achieve accurate human object tracking in complex scenarios, images from both the IR camera and EO camera were utilized. To enhance the features in these images, DL-based image fusion algorithm was developed. Image fusion combines the information from the IR and EO images to create fused images that offers improved detection and tracking capabilities and enhances the tracking results in challenging situations.

This Section presents the algorithms that are compatible with the MEIRFS hardware design for sensor fusion and multi-target tracking. In particular, the U2 Fusion [24], a unified unsupervised image fusion network, is adapted to fuse visible and infrared images and provide high-quality inputs even under adversarial environments for the downstream multi-target tracking (MTT) task.

3.2.1. Sensor fusion

Infrared cameras capture thermal radiation emitted by objects, while visible cameras capture the reflected or emitted light in the visible spectrum. Therefore, infrared cameras are useful for applications involving temperature detection, night vision, and identifying heat signatures [25,26]. Visible cameras, on the other hand, are commonly used for photography, computer vision, and surveillance in well-lit conditions. Both types of cameras serve distinct purposes and have their own specific applications based on the type of light they capture. Fusing these two modalities allows us to see the thermal characteristics of objects alongside their visual appearance, providing an enhanced scene perception and improved object detection.

Image fusion has been an active field [27,28] and many algorithms have been developed. DL-based image fusion techniques are of particular interest to MEIRFS, due to their superior performance and reduced efforts for feature engineering and fusion rules. Zhang et al. [29] provides a comprehensive review of the DL methods in different image fusion scenarios. In particular, DL for infrared and visible image fusion can be categorized into autoencoder (AE), convolutional neural network (CNN) and generative adversarial network (GAN)-based methods, according to the deep neural network architecture. Since AE is mostly used for feature extraction and image reconstruction while GAN is often unstable and difficult to train, thus we consider CNN-based methods to facilitate the multi-object tracking task. To overcome the problem of lacking universal ground-truth and no-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

reference metric, CNN-based fusion constrains the similarity between the fusion image and source images by designing loss functions. Specifically, we adapt U2Fusion [24] for the MEIRFS system, which provides a unified framework for multi-modal, multi-exposure, and multi-focal fusion. However, U2Fusion [24] did not consider image registration which is the first step towards image fusion. Due to the differences in camera parameters such as the focal length and field of view, the images may not share the same coordinate system and thus image registration is necessary to align and fuse the images. We calibrate the IR and visible cameras and compute the transformation matrix offline to reduce the online effort for image registration. After image registration, the training pipeline of U2Fusion with aligned images is shown in Figure 8.

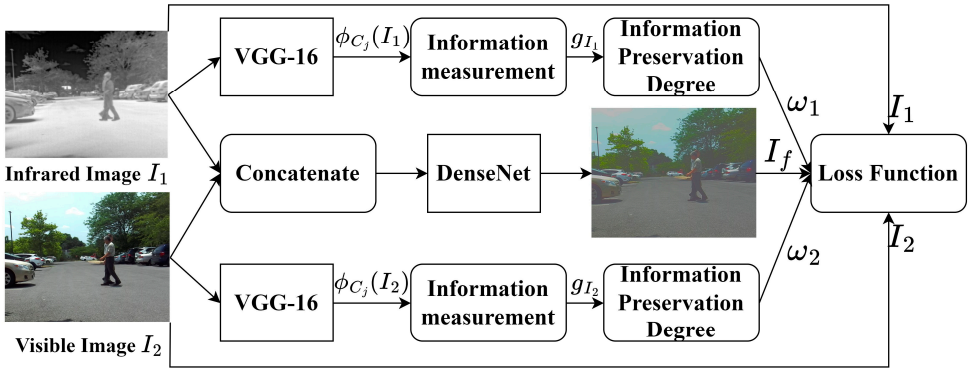

Figure 8. The training pipeline of U2Fusion [24] with aligned infrared and visible images.

To preserve the critical information of a pair of source images denoted as I_1 and I_2, U2Fusion [24] minimizes the loss function defined as follows:

$$\mathcal{L}(\theta, D) = \mathcal{L}_{sim}(\theta, D) + \lambda \mathcal{L}_{ewc}(\theta, D), \tag{1}$$

where θ denotes the parameters in DenseNet for generating the result fusion image I_f, and D is the training dataset; $\mathcal{L}_{sim}(\theta, D)$ is the similarity loss between the result and source images, $\mathcal{L}_{ewc}(\theta, D)$ is the elastic weight consolidation [30] which prevents catastrophic forgetting in continual learning, and λ is the trade-off parameter that controls the relative importance of the two parts. Additionally,

$$\mathcal{L}_{sim}(\theta, D) = \mathbb{E}[\omega_1(1 - S_{I_f,I_1}) + \omega_2(1 - S_{I_f,I_2})] + \alpha\mathbb{E}[\omega_1 \mathrm{MSE}_{I_f,I_1} + \omega_2 \mathrm{MSE}_{I_f,I_2})], \tag{2}$$

where α controls the trade-off; S_{I_f,I_i} $(i = 1,2)$ denotes the structural similarity index measure (SSIM) for constraining the structural similarity between the source image I_i and I_f while MSE_{I_f,I_i} $(i = 1,2)$ denotes the mean square error (MSE) for constraining the difference of the intensity distribution; and ω_1 and ω_2 are adaptive weights estimated based on the information measurement of the feature maps of the source images. In particular, the information measurement g_I is defined as

$$g_I = \frac{1}{5}\sum_{j=1}^{5} \frac{1}{H_j W_j D_j} \sum_{k=1}^{D_j} \left\| \nabla \phi_{C_j^k}(I) \right\|_F^2, \tag{3}$$

where $\phi_{C_j^k}(I)$ is the feature map extracted by the convolutional layer of VGG16 before the j-th max-pooling layer, and H_j, W_j, and D_j denote the height, width, and channel of the feature map, respectively. Moreover, the elastic weight consolidation $\mathcal{L}_{\mathrm{ewc}}$ is defined as

$$\mathcal{L}_{ewc}(\theta, D) = \frac{1}{2}\sum_i \mu_i(\theta - \theta^*)^2, \tag{3}$$

which penalizes the weighted squared distance between the parameter values of the current task θ and those of the previous task θ^* to prevent forgetting what has been learned from old tasks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

To train a customized model for our system, we can fine-tune the learned model in the U2Fusion using transfer learning approaches [31] with data collected by the cameras, to enhance learning efficiency. Furthermore, since sole IR or visible images can be sufficient for the object tracking task under certain environmental conditions, we design a selector switch to skip the image fusion if image fusion is unnecessary to detection the object. Figure 9 shows the complete pipeline of image fusion processing for object tracking.

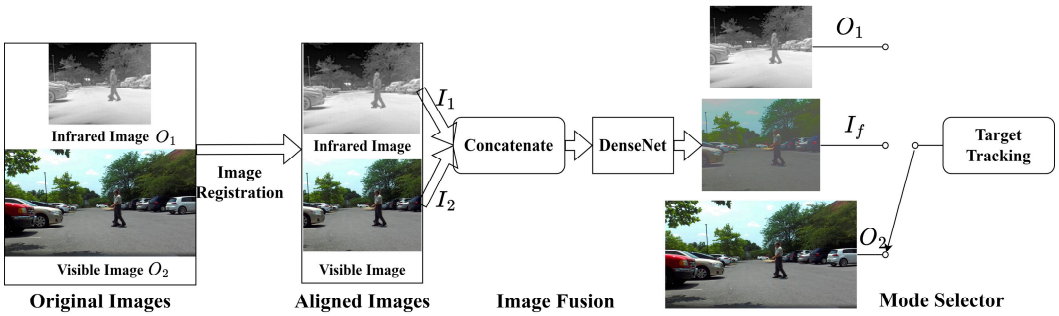

Figure 9. The pipeline of image processing for object tracking.

3.2.2. ML based algorithm for human object tracking

In certain scenarios, the human object may become lost due to inherent limitations in object detection algorithms, as well as various challenging circumstances such as occlusions and fluctuations in lighting conditions. To effectively address these situations, the utilization of a human object tracking algorithm becomes necessary [32].

To optimize the tracking results, our system employs the "ByteTrack" object tracking model as the primary algorithm [33]. For effective performance, ByteTrack utilizes YOLOX as the underlying backbone for object detection [34]. Unlike traditional methods that discard detection results below a predetermined threshold, ByteTrack takes a different approach. It associates nearly all the detected boxes by initially separating them into two categories: high-score boxes, containing detections above the threshold, and low-score boxes, encompassing detections below the threshold. The high-score boxes are first linked to existing tracklets. Subsequently, ByteTrack computes the similarity between the low-score boxes and the established tracklets, facilitating the recovery of objects that may be occluded or blurred. Consequently, remaining tracklets, which mostly correspond to background noise, are removed. The ByteTrack methodology effectively restores precise object representations while eliminating spurious background detections.

In the MEIRFS system, the fusion of IR and visible image pairs is followed by the application of the YOLOX algorithm to the fused image. This algorithm performs human object detection and generates confidence scores for the detected objects. In the presence of occlusion, priority is given to high-confidence detections, which are initially matched with the tracklets generated by the Kalman filter. Subsequently, an intersection over union (IoU) similarity calculation is utilized to evaluate the remaining tracklets and low-confidence detections. This process facilitates the matching of low-confidence detections with tracklets, enabling the system to effectively handle occlusion scenarios.

4. Experiments and results

With the integrated sensors MEIRFS system, multiple ground tests have been performed in different environments to validate the performance of each individual component in the sensor system, as well as the whole system's performance for human object detection, geolocation, and LOS friendly human object recognition.

4.1. Indoor Experiments

In Figure 10a, we tested the MEIRFS sensor system's capability of detecting and continuously tracking a single human object. When the human object appears in the IR camera's field of view, it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

will be immediately identified (marked with the red bounding box) and tracked by the sensor's system.

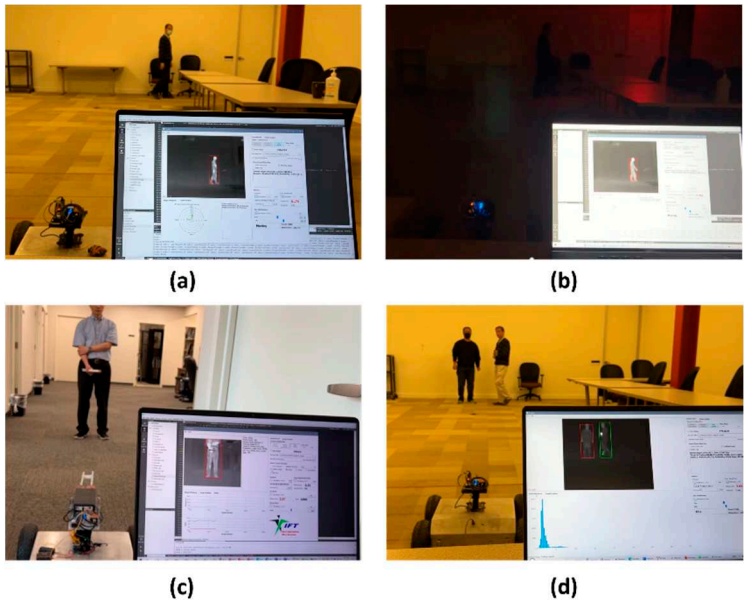

Figure 10. Experiments to demonstrate the capability of the MEIRFS sensor system.

Comparing with the traditional EO camera, one advantage of the IR camera is that it can detect the human object when there is no illumination. The long wavelength Infrared (LWIR) camera detects the direct thermal energy emitted from the human body. Figure 10b shows that the MEIRFS system can function correctly even in the dark environment.

Figure 10c demonstrates the measurement accuracy of the radar subsystem. When the friendly human object is detected by the MEIRFS system, the distance to the platform is measured by both radar subsystem and the laser rangefinder. The measurement results verified that the radar subsystem can provide accurate distance information of the friendly object, with the error of less than 0.3 meters when comparing with the laser rangefinder.

In the last test, as shown in Figure 10d, there are two human objects. The one holding the IR emitter (a heat source) is the friendly object. The other is the non-friendly object. The system was configured to tracking only the non-friendly object. When both objects came into the IR camera's FOV, the sensor system immediately identified them, and marked the friendly object with green bounding box and the non-friendly object with red box. Moreover, the sensor system immediately started to continuously track and follow the non-friendly object.

4.2. Outdoor Experiments

Extensive experiments were conducted to thoroughly validate the effectiveness of the MEIRFS system for multiple human objects tracking in outdoor environments. These experiments were designed to assess the system's performance and capabilities across various scenarios and conditions encountered in real-world outdoor settings.

The tracking model employed has undergone pre-training on two datasets, namely CrowdHuman [35] and MOT20 [36]. The CrowdHuman dataset is characterized by its extensive size, rich annotations, and substantial diversity. The CrowdHuman dataset encompasses a total of 470,000 human instances from both the training and validation subsets. Notably, each image within the dataset contains an average of 22.6 persons, thereby exhibiting a wide range of occlusions. On the other hand, the MOT20 dataset comprises eight sequences extracted from three densely populated scenes, where the number of persons per frame can reach up to 246 individuals. The pretrained model's exposure to such varied and challenging conditions enables it to effectively handle a wide

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0664.v1

array of real-world scenarios, leading to enhanced object tracking capabilities and more reliable results.

Figure 11 presents the evaluation of MEIRFS' tracking ability, revealing noteworthy insights from the top and bottom rows of the displayed results. In these scenarios, which involve the movement of multiple individuals amidst occlusion, the MEIRFS multimodal U2Fusion tracking algorithm exhibits exceptional performance. Each individual is tracked using a distinct color, showcasing the algorithm's ability to accurately track different people without experiencing any instances of object loss. The outcome underscores the robustness and reliability of the REIRFS tracking algorithm, particularly in challenging conditions where occlusion and the simultaneous presence of multiple objects present significant tracking difficulties.

Figure 11. Experiments to demonstrate the capability of the MEIRFS sensor system.

Figure 12 illustrates the performance of the MEIRFS tracking algorithm on images captured by an IR camera, images captured by a visible camera, and the fused images obtained by sensor fusion. Analysis of the top and middle rows reveals that both scenarios encounter challenges in tracking person #1, and person #2 is incorrectly assigned as person #1, while person #1 is mistakenly considered as a new individual, person #3. However, in the bottom row, following the fusion of IR and visible images, our tracking algorithm successfully tracks both person #1 and person #2, even in the presence of occlusions. The performance highlights the effectiveness of the induced sensor fusion, which combines information from both IR and visible images. As a result, the fusion process enriches the image features available for utilization by the tracking algorithm, leading to improved tracking performance in challenging scenarios.

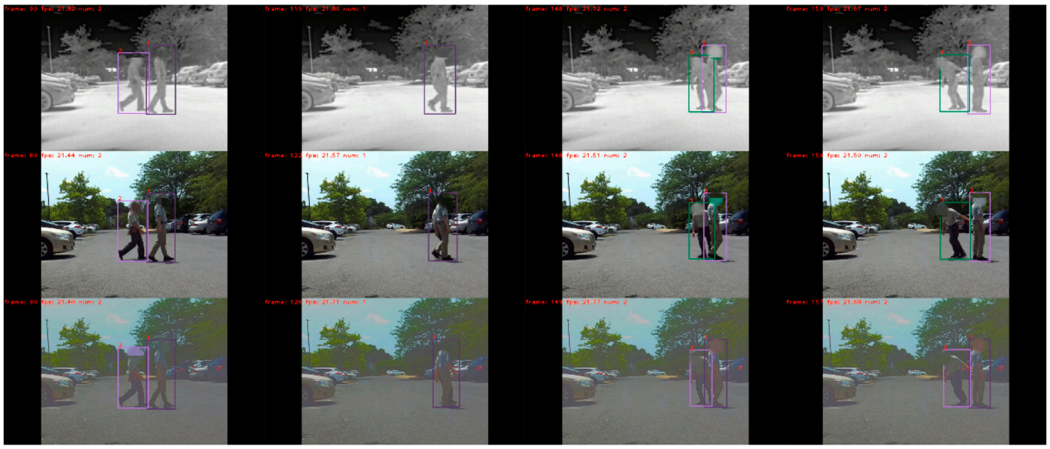

Figure 12. Results of the MEIRFS sensor system after applying sensor fusion. The top row shows the results on pictures captured by IR camera, the middle row shows the results on pictures captured by visible camera, and bottom row shows the results on fused pictures.

5. Conclusions

This paper proposes and develops a multimodal EO/IR and RF based sensor system for real-time human object detection, recognition, and tracking on autonomous vehicles. The integration of hardware and software components of the MEIRFS system was successfully accomplished and demonstrated in indoor and outdoor scenes with collected and common data sets. Prior to integration, thorough device functionality testing established communication between each device and the host computer. To enhance human object recognition and tracking (HORT), multimodal deep learning techniques were designed. Specifically, the "U2Fusion" sensor fusion algorithm and the "ByteTrack" object tracking model were utilized. These approaches significantly improved the performance of human object tracking, particularly in complex scenarios. Multiple ground tests were conducted to verify the consistent detection and recognition of human objects in various environments. The compact size and light weight of the MEIRFS system make it suitable for deployment on UGVs and UAVs, enabling real-time HORT tasks.

Future work includes deploying and testing the MEIRFS system on UAV platforms. Additionally, we aim to leverage the experience gained from ground tests to retrain the deep learning models using new images acquired from the EO/IR camera and a radar on the UAV. We anticipate that the MEIRFS system will be capable of performing the same tasks of human object detection, recognition, and tracking that have been validated during the ground tests.

References

1. Perera, F., Al-Naji, A, Law, Y., and Chahl, J. Human Detection and Motion Analysis from a Quadrotor UAV. *IOP Conf. Ser.: Mater. Sci. Eng.*, 2018; 405, 012003.
2. Rudol, P., and Doherty, P. Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery. *Aerospace Conference, IEEE*, 2008; pp. 1–8.
3. Andriluka, M., Schnitzspan, P., Meyer, J., Kohlbrecher, S., Petersen, K., von Stryk, O., Roth, S., and Schiele, B. Vision based victim detection from unmanned aerial vehicles. *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, 2010; pp. 1740–1747.
4. Gay, C., Horowitz, B., Elshaw, J.J., Bobko, P. and Kim, I., Operator suspicion and human-machine team performance under mission scenarios of unmanned ground vehicle operation. *IEEE Access*, 2019; 7, pp.36371-36379.
5. Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv*:2004.10934.
6. Chen, N., Chen, Y., et al. Enabling Smart Urban Surveillance at The Edge. *IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 109-11. 2017.
7. Munir, A., Kwon, J., Lee, J-H., Kong, J., *et al*., "FogSurv: A Fog-Assisted Architecture for Urban Surveillance Using Artificial Intelligence and Data Fusion," in *IEEE Access*, vol. 9, pp. 111938-111959, 2021.
8. Blasch, E., Pham, T., Chong, C-Y., Koch, W., Leung, H., et al. Machine Learning/Artificial Intelligence for Sensor Data Fusion–Opportunities and Challenges. *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 80-93, 2021.
9. He, Y., Deng, B., Wang, H., Cheng, L., Zhou, K., Cai, S. and Ciampa, F. Infrared machine vision and infrared thermography with deep learning: A review. *Infrared physics & technology*, 2021; 116, p.103754.
10. Huang, W., Zhang, Z., Li, W. and Tian, J. Moving object tracking based on millimeter-wave radar and vision sensor. *Journal of Applied Science and Engineering*, 2018; 21(4), pp.609-614.
11. Van Eeden, W. D., de Villiers, J.P., Berndt, R.J., Nel, W. A. J. Micro-Doppler radar classification of humans and animals in an operational environment. *Expert Systems With Applications*, Vol. 102, 1-11, 2018.
12. Majumder, U., Blasch, E., Garren, D. *Deep Learning for Radar and Communications Automatic Target Recognition*, Artech House, 2020.
13. Premebida, C., Ludwig, O. and Nunes, U. LIDAR and vision-based pedestrian detection system. *Journal of Field Robotics*, 2009; 26(9), pp.696-711.
14. Duan, Y., Irvine, J. M., Chen, H-M., G. Chen, et al. Feasibility of an Interpretability Metric for LIDAR Data. *Proc SPIE 10645*, 2018.
15. Salehi, B., Reus-Muns, G., Roy, D., Wang, Z. Jian, T., *et al*. Deep Learning on Multimodal Sensor Data at the Wireless Edge for Vehicular Network. *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639-7655, 2022.
16. Sun, S. and Zhang, Y.D. 4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach. *IEEE Journal of Selected Topics in Signal Processing*, 2021; 15(4), pp.879-891.

17. Roy, D. Li, Y., Jian, T., Tian, P., Chowdhury, K., Ioannidis, S. Multi-Modality Sensing and Data Fusion for Multi-Vehicle Detection. *IEEE Transactions on Multimedia*, vol. 25, pp. 2280-2295, 2023.
18. Vakil, A., Liu, J., Zulch, P., Blasch, E., Ewing, R., Li, J. A Survey of Multimodal Sensor Fusion for Passive RF and EO information Integration. *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 44-61, 2021.
19. Vakil, A., Blasch, E., Ewing, R., Li, J. Finding Explanations in AI Fusion of Electro-Optical/Passive Radio-Frequency Data. *Sensors* 2023, *23*(3), 1489; https://doi.org/10.3390/s23031489
20. Liu, J., Ewing, R., Blasch, E., Li, J. Synthesis of Passive Human Radio Frequency Signatures via Generative Adversarial Network. *IEEE Aerospace Conference,* 2021.
21. Liu, J., Mu, H., Vakil, A., Ewing, R., Shen, X., et al. Human Occupancy Detection via Passive Cognitive Radio. 20, 4248; *Sensors,* 2020. doi:10.3390/s20154248
22. Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z. and Qu, R. A survey of deep learning-based object detection. *IEEE access*, 2019; 7, pp.128837-128868.
23. Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M. and Lee, B. A survey of modern deep learning based object detection models. *Digital Signal Processing*, 2022; p.103514.
24. Xu, H., Ma, J., Jiang, J., Guo, X. and Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*2020; 44, no. 1, pp.502-518.
25. Liu, S., Gao, M., John, V., Liu, Z., et al. Deep Learning Thermal Image Translation for Night Vision Perception. ACM *Transactions on Intelligent Systems and Technology*, 12(1): 1-18, 2020.
26. Liu, S., Liu, H., John, V., Liu, Z., et al. Enhanced Situation Awareness through CNN-based Deep MultiModal Image Fusion, *Optical Engineering*, 59(5): 053103, 2020.
27. Zheng, Y., Blasch, E., Liu, Z. *Multispectral Image Fusion and Colorization.* SPIE Press, 2018.
28. Kaur, H., Koundal, D. and Kadyan, V. Image fusion techniques: a survey. *Archives of computational methods in Engineering,* 2021; 28, pp.4425-4447.
29. Zhang, H., Xu, H., Tian, X., Jiang, J. and Ma, J. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 2021; 76, pp.323-336.
30. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. and Hassabis, D. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017; 114(13), pp.3521-3526.
31. Bao, Y., Li, Y., Huang, S.L., Zhang, L., Zheng, L., Zamir, A. and Guibas, L. An information-theoretic approach to transferability in task transfer learning. *In 2019 IEEE international conference on image processing (ICIP)*, 2019; pp.2309-2313.
32. Xiong, Z., Wang, C., Li, Y., Luo, Y. and Cao, Y. Swin-pose: Swin transformer based human pose estimation. *In 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*2022; pp.228-233.
33. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W. and Wang, X. Bytetrack: Multi-object tracking by associating every detection box. *In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, 2022; pp.1-21.
34. Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint*, 2021; arXiv:2107.08430.
35. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X. and Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint*, 2018; arXiv:1805.00123.
36. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. and Leal-Taixé, L. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint*, 2020; arXiv:2003.09003.