Enhance Public Safety Surveillance in Smart Cities by Fusing Optical and Thermal Cameras

Nihal Poredi[a], Yu Chen[a], Xiaohua Li[a], Erik Blasch[b]
[a]Dept. of Electrical & Computer Engineering, Binghamton University, SUNY, Binghamton, NY 13902, USA
[b]The U.S. Air Force Research Laboratory, Rome, NY 13441, USA
{nporedi1, ychen, xli}@binghamton.edu, erik.blasch@us.af.mil

***Abstract*—The recent advancements in the Internet of Video Things (IoVT) and Edge-Fog-Cloud Computing paradigm make smart public safety surveillance (SPSS) a realistic solution for an effective public safety service in smart cities. Typically, a fully functional SPSS system requires multiple sensory inputs for situational awareness (SAW). As an essential component in the context of highly complex, dynamic, and heterogeneous smart city operations, SPSS is expected to be environment-resilient. Personal safety is among the top concerns of the residents in smart cities, and correspondingly pedestrian detectors are critical. Contemporary pedestrian detectors use optical cameras, whose accuracy is diminished in low-light environments, and they are rendered ineffective when obstacles block the direct line of sight to the camera. Complementary imaging sensors such as infrared have shown promise. This paper presents a full-spectrum, environment-resilient surveillance platform as an ultimate solution, which consists of multiple imaging units to cover a wide sensing spectrum. The initial hybrid pedestrian detection (HYPE) scheme is based on the fusion of data obtained from an IoVT network equipped with optical and thermal cameras. We demonstrate that training the YOLOv5 object detection model on a dataset of infrared images improves its accuracy in the detection of humans present in thermal surveillance images. A 41% decrease in objectness loss is achieved after transfer learning is performed.**

***Index Terms*—Smart Public Safety, Human Object Detection, Hybrid Thermal-Optical Cameras, Deep Learning, Data Fusion.**

I. Introduction

The Edge-Fog-Cloud Computing paradigm and Internet of Things (IoT) technology make the concept of Smart Cities become realistic, greatly improving the citizen's quality of life for a sustainable urban environment [18], [42]. Public safety service is among the most popular areas in smart city development as the safety and security of either personnel or properties are fundamental needs for an enjoyable life [16], [44]. The proliferation of Internet of Video Things (IoVT) affords an effective technological measure that makes smart public safety surveillance (SPSS) a practical solution [13], [17]. Typically, a fully functional SPSS system requires multiple sensory inputs for situational awareness (SAW) [8], specifically for safety and security-related tasks like object-of-interest detection, identification, and tracking [15], [32], [33].

Besides public safety surveillance, there is also an increasing demand for effective, efficient, and reliable surveillance solutions to maintain SAW in many mission-critical delay-sensitive tasks, such as battlefield monitoring, disaster monitoring and recovery, etc. [2], [27]. Nowadays, the optical video surveillance system is the most popular approach [12], [22]. However, it suffers from changing environmental conditions and monitoring at night, in foggy weather, on rainy days, or with wall blockages is a challenging task [39], [41]. As an essential component in the context of highly complex, dynamic, and heterogeneous smart city operations, SPSS is expected to be environmentally resilient and adaptive [14].

Personal safety is of the highest priority to the residents in smart cities, and correspondingly the capability of accurately detecting and identifying people is indispensable for safety surveillance. While it is recognized as an important component of SPSS systems, presently most contemporary pedestrian detectors use optical cameras which have a diminished accuracy in low-light environments, and they are rendered ineffective when obstacles block the direct line of sight to the camera [15]. Therefore, taking advantage of multiple sensors and leveraging modern information fusion technology are considered to address the constraints of optical cameras [5], [11], [47]. One of the candidates is thermal cameras, which are unaffected by such conditions and could be complementary.

The **FU**ll-**S**pectrum, **E**nvironment-Resilient **S**urveillance (FUSERS) framework consists of multiple imaging units to cover a wide sensing spectrum, including 5G-mmWave imaging technology, sub-6 GHz RF communication, infrared thermal imaging units, and popular optical surveillance [33]. To support full-spectrum ubiquitous surveillance studies, the FUSERS platform is expected to address the challenges that today's optical video surveillance systems is facing. The fusion of sub-6 GHz 5G signals, mmWave 5G signals and optical images will provide continuous tracking in all environments and thus realize environmentally resilient surveillance.

This paper introduces the FUSERS platform with a conceptual-level illustration of the design rationale and architecture. Then, as a case study and preliminary design, a **hy**brid **pe**destrian detection (HYPE) scheme is presented based on the fusion of data obtained from an IoVT network consisting of optical and thermal cameras. Specifically, the HYPE system is facilitated by a multi-spectral pedestrian detector which provides accurate data from both the visible and infrared regions of the electromagnetic spectrum. In normal operations with sufficient light or good vision without obstacles, videos from the optical image sensors are applied. In low-light situations or environments with obstacles, a shift from the visible spectrum to the infrared spectrum is triggered. Both RGB and thermal datasets were selected to train the You Only Look Once (YOLO) model [36] used in the multi-spectral detector. The experimental results demonstrate that training the YOLOv5 object detection model on a dataset of infrared images improves its accuracy in the detection of human objects present in

978-1-6654-7189-3/22/$31.00 © 2023 IEEE

thermal surveillance images. The main contributions of this paper include:

- A **FU**ll-**S**pectrum, **E**nvironment-Resilient **S**urveillance (FUSERS) framework is introduced, which provides a high-level information fusion (HLIF) smart surveillance system [6] that is capable of meeting the diverse demands from various delay-sensitive, mission-critical applications;
- A preliminary **hy**brid **pe**destrian detection (HYPE) scheme is designed that integrates optical cameras and thermal cameras to investigate the effectiveness of a hybrid surveillance system; and
- A proof-of-concept prototype of HYPE is constructed and an extensive experimental study is conducted that validates the design rationale and the experimental results are encouraging.

The remainder of this paper is organized as follows. Section II reviews background knowledge of object detection in the visible and infrared regions. Section III introduces the rationale and basic design of the full spectrum, environment-resilient surveillance platform, and Section IV presents the architecture of our HYPE scheme including key components and features. The prototype implementation and evaluation are discussed in Section V. Finally, Section VI concludes this paper with ongoing efforts and future directions.

II. Background and Related Work

A. Object Detection in the Visible Region

Object detection in the visible region includes a century of development including Neural Networks [38] and Support Vector Machines (SVM) [35] to detect human faces. For example, local image features were suggested to develop an object recognition system [30]. The following decade witnessed significant progress in exploiting local image features for object recognition rather than detection, which demands a higher level of performance. In more recent years, the evolution of Deep Neural Networks (DNN), particularly Convolutional Neural Networks (CNN), has spurred the development of state-of-the-art object detection [31], [34].

Krizhevsky and colleagues proposed a Deep CNN-based approach to image classification using the ImageNet dataset and achieved a record low error rate of 15.7% in 2012 [26]. This led to the Region based CNN method [21] in 2014 that offered a record 30% performance improvement over the previous best detector. Large RGB datasets such as Imagenet and MS COCO [28] were developed, which allowed researchers to have a diverse training pool that improved the accuracy of object detection and gave them a broad span of classification.

In 2016, the You Look Only Once (YOLO) object detection network was proposed [36]. A novel approach is employed in which object detection is framed as a regression problem instead of using classifiers for detection. The real-time processing capabilities made YOLO the most popular and widely adopted object detector of its time. The YOLO scheme was further improved upon in 2017, YOLO9000, which worked on 9000 object classes in real-time [37]. It was then followed by several other versions and variations, each having incremental accuracy and performance.

YOLOv5 is a state-of-the-art object detector belonging to the YOLO family. It improves upon its predecessors in terms of performance by using features such as panoptic segmentation and object tracking. Moreover, it has various lightweight versions that are suitable for IoT environments like autonomous vehicles, surveillance, and imaging. Our HYPE scheme adopts the YOLOv5 model because of the specific designs for fast computations in resource-strained environments. However, it is worth noting that all of the object detectors mentioned above are restricted to RGB data and show dips in performance when being applied in the Infrared region.

B. Object Detection in the Infrared Region

Infrared object detection focuses on establishing distinctions among the objects, their background, and their noises. Since thermal images typically lack the high resolution of RGB images, they are corrupted with noises that need to be suppressed. In addition to noises, the background also needs suppression for the object to be identified accurately [7]. IR automatic target recognition (ATR) techniques include temporal noise suppression [29] and denoising using Shearlet Transform and histogram thresholding [3]. More recent efforts led to advanced approaches such as Haar cascade classifier-based detection [40], multiscale CNNs [45], and lightweight CNNs [19] used for medical imaging, body temperature scanners, and autonomous cars. While all these methods were effective to an extent, they suffered from high complexity and are not affordable to be deployed in IoT environments with low resource availability. Moreover, thermal datasets are not as readily available as RGB datasets, which makes the rate of development slow.

III. Full-Spectrum Environment-Resilient Surveillance Platform

Nowadays, the most popular optical video surveillance systems suffer from changing environmental conditions [39], [41]. Also, they are unable to detect weapons concealed under clothes, or people behind walls. The coverage is limited to important areas only due to both cost and privacy concerns. Theoretical models have been developed such as the National Imagery Interpretability Rating Scales (NIIRS) standard [4], [24]; however, these models do not include attributes of data trust (e.g., machine learning), security (e.g., blockchain), and bandwidth (e.g., 5G).

Due to its ubiquitous coverage [23], [25], the fifth generation and beyond cellular system (5G/BCS) can play an important role in SAW well beyond simply providing connectivity support for existing surveillance systems. As 5G signal bandwidth becomes high and the network becomes dense, the 5G communication signals potentially can be used directly for surveillance sensing and imaging purpose. Especially, the 5G millimeter-wave (mmWave) signals have enough bandwidth to produce high-resolution surveillance images similar to the Terahertz (THz) or X-ray-based imaging systems used in airports [1]. Existing THz or X-ray-based systems are either too expensive or physically too big for ubiquitous deployment. Conventional surveillance systems, such as optical cameras and radars, provide coverage over important areas only. In contrast, 5G coverage is ubiquitous and 5G nodes can be lightweight and

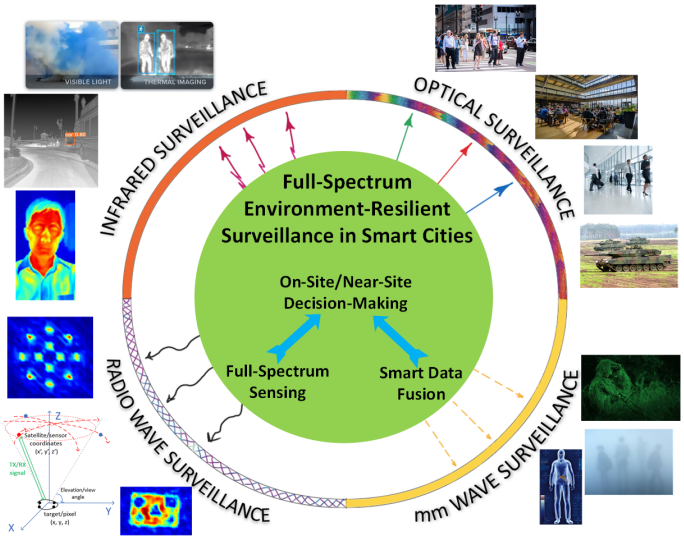

Fig. 1. A conceptual illustration of our FUSERS platform in smart cities.

low-cost. 5G network densification will make 5G nodes close to targets. As a result, 5G signals may provide a full-spectrum ubiquitous sensing mechanism to compensate for the shortness of conventional means.

The **FU**ll-**S**pectrum, **E**nvironment-Resilient **S**urveillance (FUSERS) platform consists of multiple imaging units to cover a wide sensing spectrum, including 5G-mmWave imaging technology, sub-6 GHz RF communication, infrared thermal imaging units, and popular optical surveillance [33]. Meanwhile, the use of many high-quality surveillance devices also introduces the challenges of big data. It is non-trivial for the dynamic data-driven adaptive computations to effectively process the exponentially increasing data volume for advanced surveillance tasks such as simultaneous target identification, tracking, and behavior analysis/prediction [10]. In addition, the ubiquitously deployed sensors make it not practical to depend on a central cloud facility. Intelligence at the edge is required for instant suspicious object identification and early alert generation. Furthermore, context understanding, which is critical for situation awareness, is established based on coordination between users and devices [9], [43]. Therefore, a human-in-the-loop, live-video computing architecture is compulsory.

Figure 1 is a conceptual level illustration of our FUSER system, which integrates 5G-mmWave, Optical, and Sub-6 GHz RF together and architecturally consists of a cluster of heterogeneous surveillance and computing devices, from THz transceivers to optic video to RF. They are integrated as edge computing cells to process the collected data on-site. The edge computing paradigm relieves the burden on communication networks and enables real-time data processing and instant on-site or near-site decision-making. These advanced features make FUSER platform the first of its kind.

Environment resilient SPSS systems also need multiple types of sensing units to function at its full potential. As shown in Fig. 1, FUSERS is a framework in which information from the visible, infrared, microwave (specifically mmWave) and radio wave regions is utilized concurrently to make decisions based on a more comprehensive view of the smart city environment. Each area of the spectrum corresponds to particular types of

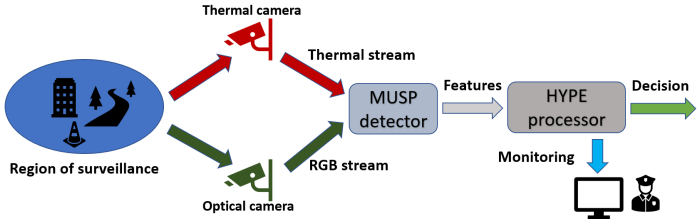

Fig. 2. HYPE System Overview.

IoT devices present in a smart city. For instance, thermal detectors would work better in conditions of low light, or obstructions blocking the direct line of sight to a security camera. A FUSERS-enabled security system will be able to correctly detect pedestrians in both thermal and RGB video streams captured by infrared and optical cameras respectively. It will also transmit information such as number of objects, coordinate information, and timestamps to the on-site or near-site decision-making unit. Better data provided to the processor will ultimately lead to better decisions. The combination of data from multiple bands to fill up the blind spots and make well informed decisions is what makes the FUSERS platform environment resilient.

IV. HYPE System Architecture

As the initial step toward a complete FUSERS system design, the system integrates two different imaging sensors, the normal optical cameras and infrared thermal cameras in an IoVT surveillance network. Figure 2 presents the hybrid pedestrian detection (HYPE) architecture and the working flow of the HYPE-enabled SPSS system. The data processing block of the system is the decision-making unit. Improving the quality of data provided to HYPE will naturally bring about an elevation in performance.

HYPE leverages a **MU**lti-**S**pectral **P**edestrian (MUSP) detector that provides accurate data from both the visible and infrared regions of the electromagnetic spectrum. The MUSP detector receives two video streams from optical and thermal cameras respectively. The cameras are installed to observe a certain region of surveillance; typically a street intersection, or a high-security alley. The MUSP detector extracts various features such as the number of pedestrians, averaged coordinate information, and timestamps from the video streams. The feature extraction uses information from both regions of the spectrum. The EO and IR features are fused and then passed on to the data processing block for decision-making. The diversification of data used in the decision-making process is what contributes to the enhanced accuracy of the SPSS system.

A MUSP detector is able to perform detection in multiple regions of the electromagnetic spectrum. The proposed MUSP detector identifies pedestrians appearing in both RGB and thermal video streams. Traditional EO/IR cameras offer the ability for simultaneous images, co-referenced collections, and similar operational conditions to support object detection, recognition, classification, and identification.

The effectiveness and efficiency of the MUSP detector depend on the quality of the feature fusion algorithm, which is based on transfer learning. Due to the much less thermal imagery available than the abundant RGB dataset in public

domains, there are fewer state-of-the-art thermal object detectors as compared to their RGB counterparts. Moreover, the lightweight nature of recent RGB object detectors, makes them ideal for IoT environments like SPSS systems, where computational resources are constrained. Therefore, it is reasonable to consider transfer learning technology to train an RGB detector on thermal images to produce a multi-spectral detector.

The base model adopted in this paper is YOLOv5, which is one of the most recent and most accurate object detector trained on the MS COCO dataset which is in the RGB format [28]. The YOLO version used is the YOLOv5s, a lightweight model that is suitable for IoVT network at the edge. The model is then trained on thermal images extracted from the FLIR dataset [20], which contains thermal images in the grayscale format and their respective labels in the form of txt files. Using the labeled data enables the model to learn features that are specific to pedestrians in both the visible and infrared regions of the spectrum. Ultimately, the fusion of EO and IR data leads to better accuracy of pedestrian detection in thermal streams.

V. Experimental Results

A. Dataset Description

Both RGB and thermal datasets were used to train the model used in the MUSP detector. The YOLOv5s model was trained on the COCO dataset containing 330,000 RGB images, 91 classes, and 2.5 million labels. This comprehensive dataset allows the model to perform widespread classification of objects, making it an ideal candidate for transfer learning applications. In this work, transfer learning is done using the FLIR Thermal Images dataset [46]. It consists of over 20,000 thermal and RGB images captured from the dashboard camera of a car driving around the streets of a city in various light conditions. As such, some images contain pedestrians while others do not. The bounding box regions marking the pedestrians are specified in the annotations, which were processed to generate labels. The labels were in the form of individual '.txt' files corresponding to each image as per the YOLOv5 format. 80% of the images were used for training, 15% for validation, and 5% were used for testing.

B. Model Training

The base YOLOv5 model was trained on 20,000 thermal images with their corresponding labels in the YOLOv5 format. All the layers of the model were trained and none of them were frozen, allowing for better accuracy for pedestrian detection in the thermal region. The model was trained for a total of 35 epochs in 2 phases of 25 and 10. As shown in Figure. 3, the bounding box loss (measure of bounding box accuracy) and the objectness loss (measure of prediction confidence) during training and validation, decreased over the first 25 epochs and continued to reduce marginally over the last 10 iterations. The results from Fig. 3 indicate that 25 is a sufficient number of epochs needed to achieve reasonable results. Since the training and validation images are all thermal in nature, it is assumed that the system would support detecting humans in the thermal region. Therefore, it is fair to infer that the final model, obtained through transfer learning, that it is better tuned to detect pedestrians in thermal images when compared to the base model.

C. Detection Accuracy

The detection accuracy is measured by the objectness loss and is further understood by analyzing the testing dataset results. The testing dataset consists of 1000 images selected from the FLIR dataset, along with their respective labels. Both the base model and the trained model were evaluated on the same testing dataset to maintain consistency. Figure 4 depicts some of the results obtained. The first row represents the detection results of the base model while the second row contains those of the trained model. The first column presents multiple pedestrians farther in the frame which was undetected by the base model but was correctly detected by the trained model. The second column shows that, unlike the base model, the trained model is also able to detect cyclists in addition to pedestrians on foot. The third column contains pedestrians near and farther in the frame.

While the base model successfully detects closer pedestrians, it is unable to detect faraway pedestrians. The trained model has no such issue and detects all the pedestrians irrespective of their proximity to the camera. The last column is an example of a false positive, where the base model incorrectly classifies a tree top as a person, while the trained model classifies it as a true negative. As such, it is evident that the fusion of features learned through transfer learning enables the trained model to be more accurate in identifying pedestrians in thermal images and videos.

D. Precision and Recall

Precision is a measure of the ratio of true positives to the total number of positives, including false positives. Figure 5 shows that the precision after 25 epochs increases to 0.891 and has a marginal decrease to 0.88 after 10 more epochs. This is a reasonable increase from the original precision of just under 0.7.

The ground truth coordinates for the bounding boxes are provided in the labels file for each image as part of the training dataset. The predicted coordinates form the predicted bounding box. Intersection over Union (IoU) is the measure of overlap between the predicted and ground truth bounding boxes. The Average Precision (AP) is the precision for all data points having an IoU greater than a certain threshold. The Mean Average Precision (mAP) is the AP across all classes. In this case the number of classes is limited to one. Therefore the AP will be the same as the mAP. The mAP at an IoU threshold of 0.5 after 25 epochs was 0.893, and settled at 0.88 after 10 more epochs as shown in Figure. 5. This is a considerable increase from the mAP (at an IoU=0.5) of 0.625 before training.

Recall is defined as the number of true positives divided by the sum of true positives and false negatives. Recall essentially measures the correctness of the predictions. Figure 3 shows that the recall after 25 epochs was 0.8 and fell to 0.79 after 10 more epochs. This is a substantial improvement from a recall of 0.55 before training.

The Precision Recall (PR) curve depicts the tradeoff between the two measures for a given threshold of IoU. Figure 6 shows a comparison between the PR curves at an IoU of 0.5 for both the base and trained models. It is evident from the graphs, that the trained model has a higher PR score as compared to the

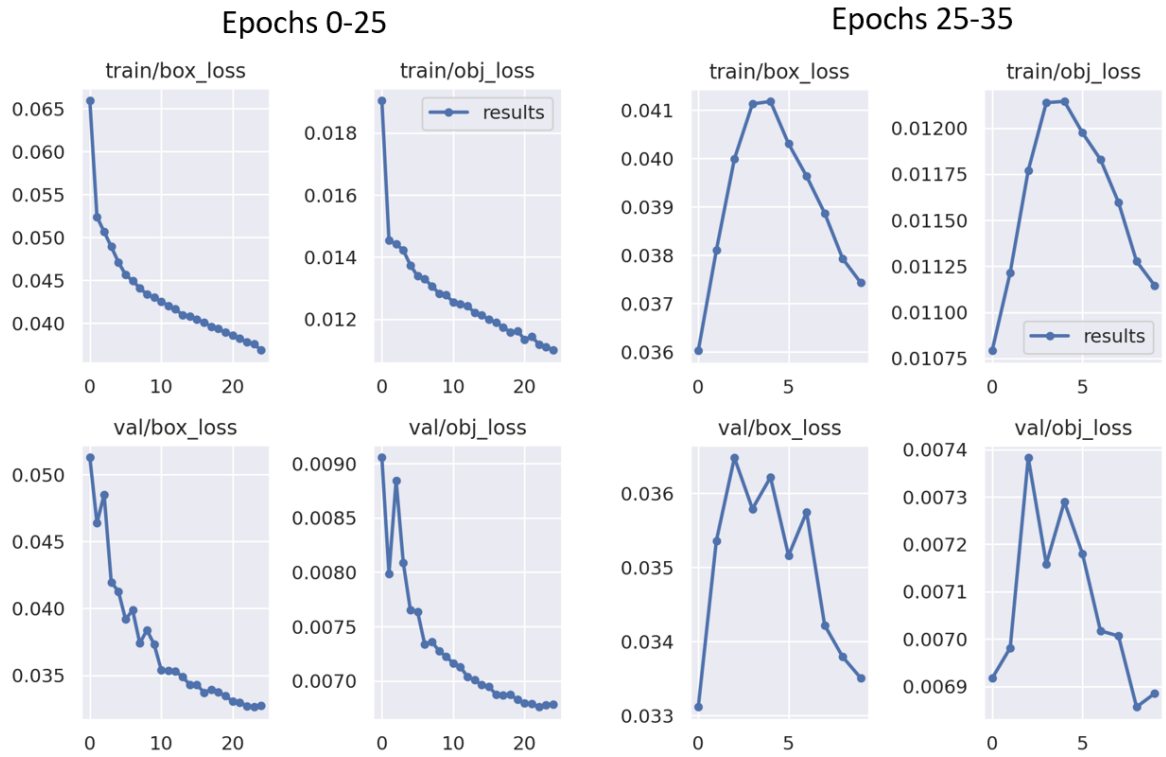

Fig. 3. Objectness and Box Loss after 25 and 35 epochs.

Fig. 4. Accuracy comparison of base and trained models.

base model. The area under the PR curve for the trained model is also larger, indicating a higher AP.

E. Discussions

It is evident from the experimental results that transfer learning enabled the model to learn features belonging to both the RGB and the thermal datasets. There is an improvement of 41% in the objectness loss and a 43% improvement in the bounding box loss. This is attributed to a reduction in the number false negatives and an increase in the number of true positives, which in turn offer reliability for surveillance operations. The ability of the detector to detect distant pedestrians is particularly encouraging from the SPSS point of view. The large and comprehensive dataset makes the training data diverse enough

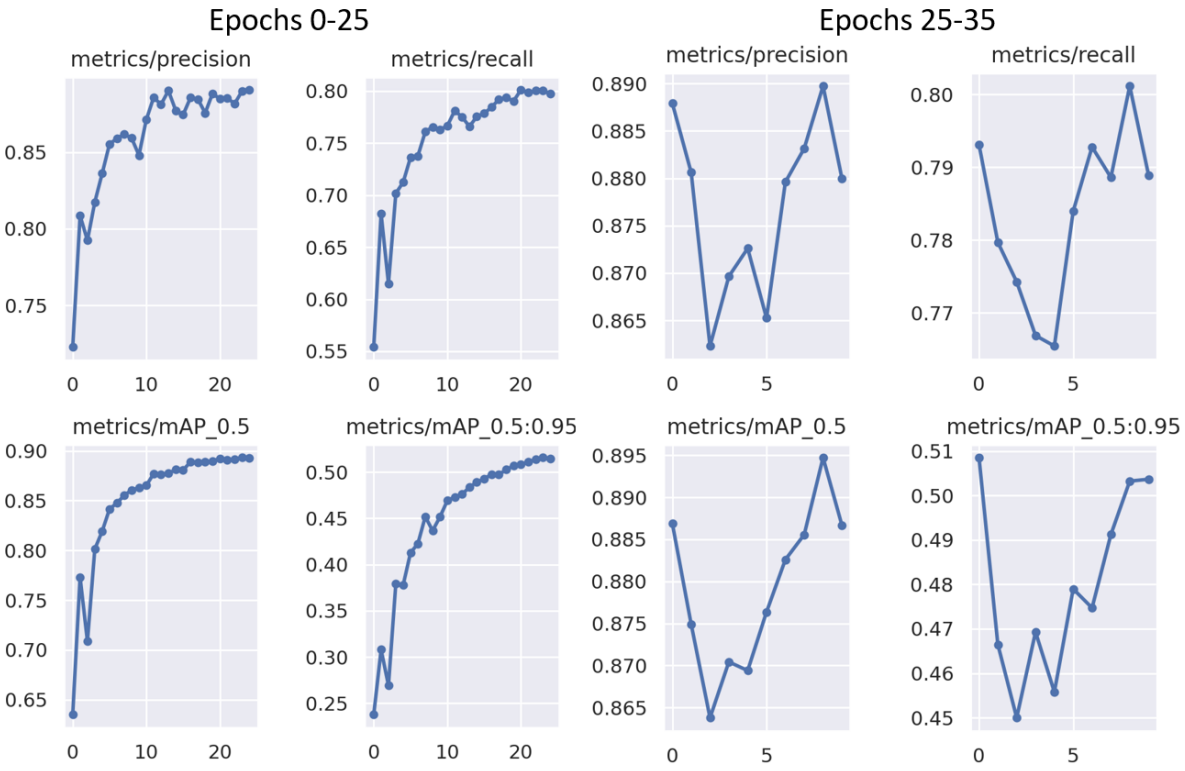

Fig. 5. Precision and Recall after 25 and 35 epochs.

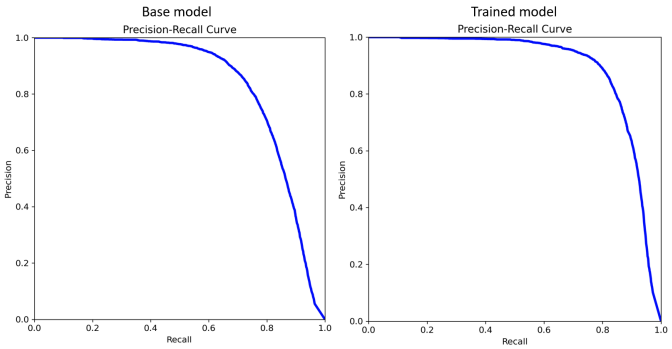

Fig. 6. Precision-Recall curves.

for real world applications. A significant improvement is not observed after 25 epochs. More clarity is needed on the ideal number of epochs relative to the size of the dataset.

VI. Conclusions and On-Going Effort

The paper introduces the full-spectrum, environmentally resilient surveillance platform, FUSERS, which integrates the 5G millimeter-wave-imaging technology with our existing sub-6 GHz RF communication testbed and optical surveillance testbed iSENSE [33]. To support full-spectrum ubiquitous surveillance studies, the platform is able to address the challenges that today's optical video surveillance systems are facing, such as interference from ever-changing environmental factors, including lighting conditions, weather changes, wall blockages, etc. The fusion of sub-6 GHz 5G signals, mmWave 5G signals and optical images will provide continuous tracking in all environments and thus realize environmentally resilient surveillance.

As an initial step, the HYPE scheme for SPSS is used as a case study. The analysis of the multi-spectral detector relative to the YOLOv5 model demonstrates its feasibility. The reasonable performance of the model in the infrared spectrum is an advantage that can be leveraged when the region of surveillance is under low light conditions. This also underscores the importance of using data from multiple sources while developing AI solutions aimed at supporting critical infrastructure. This small success verified that training the YOLOv5 object detection model on a dataset of infrared images improves its accuracy in the detection of humans present in thermal surveillance images. It is also very encouraging that there is a 41% decrease in objectness loss after transfer learning is performed.

Acknowledgement

This work is supported partially under AFRL Summer Faculty Fellowship Program (SFFP) via contracts FA8750-15-3-6003, FA9550-15-001, and FA9550-20-F-0005, and fundamental research FA9550-20-1-0237. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U. S. Air Force.

References

[1] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 219, 2015.

[2] S. S. Ahmed, A. Schiessl, F. Gumbmann, M. Tiebout, S. Methfessel, and L.-P. Schmidt, "Advanced microwave imaging," *IEEE microwave magazine*, vol. 13, no. 6, pp. 26–43, 2012.

[3] T. Anju and N. N. Raj, "Shearlet transform based image denoising using histogram thresholding," in *2016 International Conference on Communication Systems and Networks (ComNet)*. IEEE, 2016, pp. 162–166.

[4] E. Blasch, H.-M. Chen, J. M. Irvine, Z. Wang, G. Chen, J. Nagy, and S. Scott, "Prediction of compression-induced image interpretability degradation," *Optical Engineering*, vol. 57, no. 4, pp. 043 108–043 108, 2018.

[5] E. Blasch and B. Kahler, "Multiresolution eo/ir target tracking and identification," in *2005 7th International Conference on Information Fusion*, vol. 1. IEEE, 2005, pp. 8–pp.

[6] E. Blasch and D. A. Lambert, *High-level information fusion management and systems design*. Artech House, 2012.

[7] E. Blasch, Z. Liu, and Y. Zheng, "Advances in infrared image processing and exploitation using deep learning," in *Proc. of SPIE Vol*, vol. 12107, pp. 121 071M–1.

[8] E. Blasch, N. Sullivan, G. Chen, Y. Chen, D. Shen, W. Yu, and H.-M. Chen, "Data fusion information group (dfig) model meets ai+ ml," in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXXI*, vol. 12122. SPIE, 2022, pp. 162–171.

[9] E. P. Blasch and A. J. Aved, "Dynamic data-driven application system (dddas) for video surveillance user support," *Procedia Computer Science*, vol. 51, pp. 2503–2517, 2015.

[10] E. P. Blasch, F. Darema, S. Ravela, and A. J. Aved, *Handbook of Dynamic Data Driven Applications Systems: Volume 1*. Springer Nature, 2022, vol. 1.

[11] E. P. Blasch, Y. Zheng, S. Liu, and Z. Liu, "Multi-modal video fusion for context-aided tracking," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–8.

[12] G. L. Charvat, L. C. Kempel, E. J. Rothwell, C. M. Coleman, and E. L. Mokole, "A through-dielectric ultrawideband (uwb) switched-antenna-array radar imaging system," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 11, pp. 5495–5500, 2012.

[13] C. W. Chen, "Internet of video things: Next-generation iot with visual sensors," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6676–6685, 2020.

[14] N. Chen and Y. Chen, "Smart city surveillance at the network edge in the era of iot: opportunities and challenges," *Smart Cities: Development and Governance Frameworks*, pp. 153–176, 2018.

[15] N. Chen, Y. Chen, E. Blasch, H. Ling, Y. You, and X. Ye, "Enabling smart urban surveillance at the edge," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017, pp. 109–119.

[16] N. Chen, Y. Chen, X. Ye, H. Ling, S. Song, and C.-T. Huang, "Smart city surveillance in fog computing," *Advances in mobile cloud computing and big data in the 5G era*, pp. 203–226, 2017.

[17] Y. Chen, T. Zhao, P. Cheng, M. Ding, and C. W. Chen, "Joint front–edge–cloud iovt analytics: Resource-effective design and scheduling," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 941–23 953, 2022.

[18] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. A. Pardo, and H. J. Scholl, "Understanding smart cities: An integrative framework," in *2012 45th Hawaii international conference on system sciences*. IEEE, 2012, pp. 2289–2297.

[19] X. Dai, X. Yuan, and X. Wei, "Tirnet: Object detection in thermal infrared images for autonomous driving," *Applied Intelligence*, vol. 51, pp. 1244–1261, 2021.

[20] FLIR Group, "Flir thermal dataset for algorithm training," https://www.flir.com/.

[21] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[22] H. Griffiths and C. Baker, "Passive coherent location radar systems. part 1: Performance prediction," *IEE Proceedings-Radar, Sonar and Navigation*, vol. 152, no. 3, pp. 153–159, 2005.

[23] A. Hassanien, M. G. Amin, Y. D. Zhang, and F. Ahmad, "Dual-function radar-communications: Information embedding using sidelobe control and waveform diversity," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 2168–2181, 2015.

[24] B. Kahler and E. Blasch, "Predicted radar/optical feature fusion gains for target identification," in *Proceedings of the IEEE 2010 National Aerospace & Electronics Conference*. IEEE, 2010, pp. 405–412.

[25] R. J. Kozick and S. A. Kassam, "Synthetic aperture pulse-echo imaging with rectangular boundary arrays (acoustic imaging)," *IEEE Transactions on Image Processing*, vol. 2, no. 1, pp. 68–79, 1993.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[27] L. Li, J. McCann, N. Pollard, and C. Faloutsos, "Bolero: a principled technique for including bone length constraints in motion capture occlusion filling," in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2010, pp. 179–188.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[29] D. Liu and Z. Li, "Temporal noise suppression for small target detection in infrared image sequences," *Optik*, vol. 126, no. 24, pp. 4789–4795, 2015.

[30] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[31] K. L. Masita, A. N. Hasan, and T. Shongwe, "Deep learning in object detection: A review," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2020, pp. 1–11.

[32] S. Y. Nikouei, Y. Chen, A. Aved, E. Blasch, and T. R. Faughnan, "I-safe: Instant suspicious activity identification at the edge using fuzzy decision making," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 101–112.

[33] S. Y. Nikouei, Y. Chen, S. Song, B.-Y. Choi, and T. R. Faughnan, "Toward intelligent surveillance as an edge network service (isense) using lightweight detection and tracking algorithms," *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 1624–1637, 2019.

[34] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Real-time human detection as an edge service enabled by a lightweight cnn," in *2018 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 2018, pp. 125–129.

[35] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1997, pp. 130–136.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[37] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[38] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[39] M. Seifeldin, A. Saeed, A. E. Kosba, A. El-Keyi, and M. Youssef, "Nuzzer: A large-scale device-free passive localization system for wireless environments," *IEEE Transactions on Mobile Computing*, vol. 12, no. 7, pp. 1321–1334, 2012.

[40] C. H. Setjo, B. Achmad *et al.*, "Thermal image human detection using haar-cascade classifier," in *2017 7th International Annual Engineering Seminar (InAES)*. IEEE, 2017, pp. 1–6.

[41] T. Shan, S. Liu, Y. D. Zhang, M. G. Amin, R. Tao, and Y. Feng, "Efficient architecture and hardware implementation of coherent integration processor for digital video broadcast-based passive bistatic radar," *IET Radar, Sonar & Navigation*, vol. 10, no. 1, pp. 97–106, 2016.

[42] B. N. Silva, M. Khan, and K. Han, "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities," *Sustainable cities and society*, vol. 38, pp. 697–713, 2018.

[43] L. Snidaro, J. Garcia-Herrera, J. Llinas, and E. Blasch, "Context-enhanced information fusion," *Boosting Real-World Performance with Domain Knowledge*, 2016.

[44] R. Xu, S. Y. Nikouei, D. Nagothu, A. Fitwi, and Y. Chen, "Blendsps: A blockchain-enabled decentralized smart public safety system," *Smart Cities*, vol. 3, no. 3, pp. 928–951, 2020.

[45] D. Zeng and M. Zhu, "Multiscale fully convolutional network for foreground object detection in infrared videos," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 617–621, 2018.

[46] R. Zhang, J. Bin, Z. Liu, and E. Blasch, "Wggan: A wavelet-guided generative adversarial network for thermal image translation," in *Generative Adversarial Networks for Image-to-Image Translation*. Elsevier, 2021, pp. 313–327.

[47] Y. Zheng, E. Blasch, and Z. Liu, *Multispectral image fusion and colorization*. SPIE press Bellingham, Washington, 2018, vol. 481.