ABSTRACT For ordinary IR-illuminators in nighttime surveillance system, insufficient illumination may cause misdetection for faraway object while excessive illumination leads to over-exposure of nearby object. To overcome these two problems, we use the MI3 dataset as our benchmark dataset for modern object detection methods. We first provide complete annotations for the MI3 as its current ground-truth is incomplete. Then, we use these multi-intensity illuminated IR videos to evaluate several widely used object detectors, i.e., SSD, YOLO, Faster R-CNN, and Mask R-CNN, by analyzing the effective range of different illumination intensities. By including a tracking scheme, as well as developing of a new fusion method for different illumination intensities to improve the performance, the proposed approach may serve as a new benchmark of face and object detection for a wide range of distances. The new dataset (Dataset is available: https://ieee-dataport.org/documents/mi3) with more complete annotations and source codes (Codes are available: https://github.com/thesuperorange/deepMI3) is available online.

INDEX TERMS Object detection, nighttime surveillance, face tracking, multi-intensity IR illuminator.
such as YOLO [8]–[10] and SSD [11] do not have a region cropping module. They are simpler and faster than two-stage detectors, but have trailed behind in detection accuracy.

In this paper, we will consider single-stage detectors such as SSD and YOLOv4, and two-stage ones such as Faster R-CNN and Mask R-CNN, in the experiments. As different applications use infrared images in quite different ways, it is not possible to establish a universal IR dataset; therefore, credibly pretrained model of the above detectors are experimented on the MI3 dataset to setup a baseline for quantitative evaluation of the effect of adopting the multi-intensity illumination. For example, examination of confidence value of deep learning-based object detection may suggest the number of illumination intensities required for object detection for an extended range of distance. Moreover, we may also identify an effective range wherein reasonable detection results can be achieved with one or more illumination intensities of the multi-intensity IR illuminator.

Invasion detection and face recognition are two major topics in the nighttime surveillance system. Despite the effectiveness of powerful deep neural networks in object detection, poor lighting conditions often cause false detections, especially for faces. While some works combine the information from both visible and infrared images [12], [13], we focus on multi-intensity IR images in this work, under the scenario of night-time surveillance, and refine face detection results by information obtained from consecutive image frames. The approach is to combine high-quality face detection [14] and generic tracking [15] to improve both precision and recall of the face detection. To further increase the accuracy, we take advantage of complementary illumination conditions and propose a fusion method for object/face detection. Thus, main contributions of this paper include:

- A new criterion is established for evaluating different detection methods based on MIIR and analyzing the contribution of video channel of a specific illumination intensity.
- A tracking method is presented for refining face detection results to increase the F-measure of face detection.
- A fusion method is proposed to effectively merge information obtained from multiple channels to achieve higher accuracy in object/face detection.

### II. PROPOSED APPROACH

In this paper, we consider two major tasks in nighttime surveillance, i.e., invasion (human) detection and face region recognition, and examine the range of distance for them to work effectively with MIIR. While the former only need to provide the information of the existence of people in the scene, whether they are intruder or not, the latter is essential if the identification of each potential intruder is needed.

In the following subsections, we first introduce the MI3 dataset for nighttime surveillance based on MIIR. Then, we use such dataset to analyze the effective ranges of object and face detection using some baseline CNN models. As the two tasks require different image resolutions, they are considered separately in Sec. II-B. Lastly, we provide a tracking method for refining the face detection results and a fusion method for improving the detection accuracy in Sec. II-C and Sec. II-D, respectively.

#### A. MI3 DATASET OVERVIEW

The MI3 dataset [2] contains five different scenes, each having various patterns of ordinary people movements, as shown in Table 1. The dataset contains a total of 32,346 images which are separated into six sub-videos (channels), with each channel corresponding to a fixed intensity of illumination. Since the ground-truth in the original MI3 dataset is incomplete, and also out of date, we try to re-annotate human bounding boxes by following the Pascal VOC [16] format. To generate more convincing ground-truth bounding boxes, we obtained consensus labels from MTurk, before verifying and readjusting them manually. We labelled objects in all images for object detection, and labelled faces as well but only in specific scenes wherein single/multiple people are walking toward the camera (from far to near) with face regions available for detection.

#### B. EFFECTIVE RANGE ANALYSIS

First, we select some deep learning methods for both human and face detection. For human detection, we apply both single-stage detectors (SSD [11], YOLOv4 [10]) and two-stage detectors (Faster R-CNN [6], Mask R-CNN [7]), as baseline methods, wherein all methods are pre-trained on MSCOCO dataset [17]. While both Faster R-CNN and Mask R-CNN adopt ResNet-101 [18] as backbone, SSD and YOLO are established with MobileNet_v2 [19] and Darknet, respectively. As for face detection, we use Faster R-CNN object

---

**TABLE 1. Different scenes in MI3 dataset.**

| Scenes     | Video clips                        |
|------------|-----------------------------------|
| Pathway    | Pathway1, Pathway2_1, Pathway2_2, Pathway2_3 |
| Bus        | Bus2, Bus3                         |
| Doorway    | Doorway1, Doorway2, Doorway3       |
| Room       | Room1, Room2                       |
| Staircase  | Staircase1, Staircase2, Staircase3, Staircase4 |

---

1https://www.mturk.com/

2We use term ‘YOLO’ in the following section for brevity.
detector\(^3\) based on ResNet-101 but retrained on WIDER face dataset [20].

In [3], the advantage of multi-intensity illumination is demonstrated qualitatively that both far and near objects can be captured with better image quality, which is infeasible previously with a fixed illumination intensity. In this paper, quantitative evaluation of image quality of each channel is established via the definition of \textit{effective range} for different detectors, wherein reasonably high confidence scores are obtained for an individual channel, or for a number of channels. Although MI3 dataset has 6 channels, only Channels 2, 4, and 6 are considered in the following for brevity.

1) OBJECT DETECTION

For object detection, we use frames 350 to 540 of Pathway1 as an example, as shown in Fig. 1, to demonstrate the basic idea of effective range because of its simplicity. We use frame number of Pathway1 to represent “relative distance” as the scene contains a person walking from far to near in a roughly constant speed. Fig. 2 shows the trend of confidence score and IoU (intersection over union) [21] results, with the latter appeared to be more fluctuated, of object detection obtained with different CNN models for the three selected video channels. To filter out apparent outliers in the detection, simple \(3 \times 1\) median filtering is applied to each time series. Moreover, if the confidence is larger than 0.8 and continues for more than 10 consecutive frames, it is represented by a line plot; otherwise, it is represented by a scatter plot with partially transparent (lighter) color. Similar process is adopted to represent to IoU but using a threshold of 0.5.

It can be observed from the above line plots that almost every confidence scores are higher than IoUs and have smoother and more continuous (except for SSD) line plots. Thus, we use the line plot of confidence score in Fig. 2 to define the effective range (in terms of number of frames of high confidence). If there is a discontinuous problem such as Channel 2, 4 and 6 in SSD and Channel 6 in Faster R-CNN, the gap will be connected if the corresponding line plot of IoU exists. The effective range ends when the confidence has a 5\% drop in 10 consecutive frames.

In general, the confidence scores of SSD are the lowest, and result in the shortest and most fluctuated segments of line plots, which appear to represent less meaningful “effective range”. For YOLO and Mask R-CNN, the effective range of Channel 6 covers the other two, i.e., it starts later than Channels 2 but ends as same as Channel 2 and 4. The major benefit of using the multi-intensity IR videos is the complementary effect of using various illumination intensities, which is most apparent from the effective ranges of Faster R-CNN, i.e., brighter (darker) channels are more capable of detecting human in the far field (near field). Such effect also exist in similar results obtained with YOLO and Mask R-CNN, as can be examined more conveniently from a simpler illustrations of effective ranges provided in Fig. 3. As for Mask R-CNN, effective ranges of the brighter channels always cover those of darker channels except for the very last frame of the nearest foreground object.

To further examine the above complementary effect, unions of the effective range, denoted as \textit{composite range}, are also depicted in Fig. 3. Note that the unions for YOLO and Mask R-CNN almost have the same range as Channel 6 while the composite range of Faster R-CNN increases 11\% from the effective range of Channel 6 toward the near field. On the other hand, the composite range of SSD increase 14\% from the effective range of Channel 2 toward the far field. Overall, MIIR can indeed extend the depth of field (DoF) of surveillance but may have different effects for different CNN models. Nonetheless, if the object detection accuracy, instead of image quality, is the main concern, only Channels 2 and 6 are necessary for all methods.

\(^3\)The reason for selecting Faster R-CNN will be more clear later.
2) FACE DETECTION

Intuitively, human detection is mainly concerned with object contours, while face detection need to take into account the content of face region. Therefore, the following investigation for face detection adopts the Pathway2_3 video clip which contains a maximum of five people walking from far to near, and often having occlusions of their bodies, as shown in Fig. 4. Instead of considering all models, only results of the Faster R-CNN, which demonstrates high inter-channel complement and has the largest composite range of surveillance, are shown for brevity. Fig. 5 shows the face detection results for four representative people in Pathway2_3, wherein only the confidence scores are shown as the size and shape of the face region are much simpler than the human body. By following the threshold setting suggested in [22], the threshold for confidence score is 0.5 (cf. 0.8 suggested for object detection, as in Fig. 2).

Overall, the start (and end) of effective range of a particular channel shown in Fig. 5 is different for each person because of different distance to the camera. However, the major difference between Fig. 5 and Fig. 2(c) is the relationship between the coverage of the effective range and the image brightness: while brighter channels always have larger coverage of such range for in human detection, it is not always true for face detection, as the latter needs more detailed information within the face region. In general, the brightest channel can detect face earlier but it may fail to detect a face due to the overexposure problem when the face is too close to the camera, as one can observe from Figs. 4(a)-(c). Although darker channel will result in later detection of a face, face images of better quality are expected, as higher resolution of the face region can be obtained for a person closer to the camera.

On the other hand, since good image quality as well as face pose are both very important for subsequent face recognition, if any, all channels are equally indispensable for effective wide area video surveillance at night time. This is because some complex situations may impair the face detection results for different channels at different time instances, which include: (i) occlusion among different people, e.g., person 4 in frame 430 (Fig. 4(d)), and (ii) non-frontal face pose due to self-occlusion, e.g., person 1 in frame 450 (Fig. 4(e)).

C. REFINEMENT OF FACE DETECTION VIA TRACKING

To further improve the face detection results obtained for each illumination intensity, we modify the approach proposed in [23] and employ the MDnet [15] as our tracker. As shown in Fig. 6, whenever a face is detected, it becomes a target frame \( F_0 \) and initiates a new face track. The MDnet is then adopted as our tracker to track the target face for \( N \) frames (\( F_1 \) to \( F_{10} \)), with \( N = 10 \) selected in our approach, and generate tracking results \( t_{i1} \) to \( t_{10} \) as shown in Fig. 6(a). Then, the IoU comparison between the tracking result \( t_i \) and the detected results \( d_{mi} \) in the bounding box pool is performed (blue double arrow on the left), where \( t_i \) is the tracking result and \( d_{mi} \) is the \( m \)th detection result in frame \( i \), respectively.

For the IoU threshold (\( \theta_{IoU} \)) set to 0.3, the IoU count is defined by

\[
\gamma_i = \begin{cases} 
1, & \text{if } \max(IoU(t_i, d_{mi})) > \theta_{IoU} \\
0, & \text{otherwise.}
\end{cases}
\]

If \( C \geq \frac{N}{2} \), this tracklet will be added to the track (black arrow); otherwise, it will be dropped and the track will be terminated. When the tracklet is qualified to be added, the tracklet will either pick detection result or tracking result, as shown with

4 As face detection involves more complex image features, false alarms often occur for each video channel and need to be removed via tracking, before additional improvements can be obtained by decreasing miss detections via inter-channel fusion discussed in the next subsection.
FIGURE 6. Overview of the tracking approach. (a) Start from a target frame, compare the IoU between tracking results and detection results in the bounding box pool. (b) Pick tracking results (yellow) if no detection results (green) is matched.

check marks in Fig. 6(b). Specifically, an element of the tracklet \( tr_k \) is determined with

\[
tr_k = \begin{cases} 
\arg \max_i (\text{IoU}(t_k, d^i_k)) \\
0_k, & \text{if } \max_i (\text{IoU}(t_k, d^i_k)) > \theta_{\text{IoU}} \\
\text{otherwise} 
\end{cases}
\]

When a detection result is picked by a tracklet, it will be removed from the bounding box pool. Next, we take the last frame picked from detection results, e.g., \( d_9 \) in Fig. 6, as the new target and start a new tracklet. Such flow for a track will continue until a track is completed or terminated. A new track will start for the next detection of an orphan from the bounding box pool.

D. FUSION OF INTER-CHANNEL INFORMATION

In Sec. II-B, we demonstrate the advantage of multi-intensity IR-illuminator through the examination of the complementary effect among different effective ranges of different illumination intensities from the simple Pathway1 scene which only contains a single person walking from far to near. An intuitive way of using the effective range for the above simple example is to switch twice (from Channel 6 to Channel 4, and then from Channel 4 to Channel 2) from brightest channel to the darker ones as the person is moving toward the camera. For more general situations with multiple people moving arbitrarily in the scene, a slightly more complicated method is adopted in our experiments to improve the detection performance, for both far and near objects, which consists of the following steps:

1) Apply object detection to each video channel.
2) Remove object bounding boxes (BBs) with low (less than 0.8) confidence scores.
3) Identify pairs of inter-channel BBs with reasonable (larger than 0.3) IoUs.
4) Remove BBs with smaller confidence for each BB pair.
5) Consider the remaining BBs for further processing.

Fig. 7 shows an example of the above fusion process. Thus, the complementary effect among different channels will be fully explored as we will then have (i) the best representative BB for a group of nearby BBs each from a specific channel or (ii) the representative BB for an image region wherein no other illumination level can detect an object. Consequently, previously mentioned overexposure and insufficient lighting problems may be resolved after such fusion process, as shown in Fig. 8 wherein all selected BBs correspond to confidence scores higher than 98.5%.

III. EXPERIMENTAL RESULTS

In this section, we evaluate results obtained with different methods proposed in the previous section. First, we compare different detectors with respect to their effective range defined in Sec. II-B. Then the evaluations of inter-channel fusion (Sec. II-D) and face tracking results (Sec. II-C) are provided to demonstrate their effectiveness in performance improvements.

A. EFFECTIVE RANGE EVALUATION

Since we focus on frame-based detection, we follow the standard COCO-style Average Precision (AP) for object
FIGURE 9. Average APs of object detection for Pathway1 dataset based on (i) effective ranges shown in Fig. 3, and (ii) the complete range.

detection, while the AP is computed by averaging the APs for IoU = 0.5, 0.55, ..., 0.95. Fig. 9 shows the AP thus computed for each object detection method (for different channels and for confidence thresholds θ = 0.5, 0.8, 0.9 and 0.98), wherein regular color corresponds to results obtained for effective range for the simple example shown in Fig. 3 while light color corresponds to the usage of complete range. It is readily observable that the detection can be improved significantly from the results obtained for the complete video clip if only effective range of each channel, as shown in Fig. 3, are taken into account.

As for the comparison among different detectors, SSD has significant drops for increasing IoU thresholds while others only have significant drops for IoU threshold changing from 0.9 to 0.98. Moreover, one can see from results obtained for effective ranges that YOLO has best detection results for the highest illumination level (Channel 6), while Faster R-CNN has best results for Channels 2 and 4. (For Channel 4, Mask R-CNN actually has detection results comparable to Faster R-CNN for thresholds less than 0.98.)

B. EVALUATION OF INTER-CHANNEL FUSION

For the evaluation of the simple information fusion scheme described in Sec. II-D, object detection results are considered in this subsection for individual channels, as well as combinations of multiple channels. Table 2 presents object detection results by considering (i) Channels 2, (ii) Channels 4, (iii) Channels 6, (iv) two channels, and (v) all channels. For (iv), Channels 2 and 6 are selected as the union of their effective ranges, as shown in Fig. 3, do cover the same range as with (v). The best results of each detection scheme are marked in boldface for each of the five scenes shown in Table 1. Overall we have total numbers of best results equal to 4, 2, 3, 16 for (i), (ii), ..., and (v), respectively. Such results show that the complementary effect is actually among all illumination intensities, as with (v), instead of merely depending on the coverage of effective ranges, which is the same for both (iv) and (v). Thus, it is suggested that all channels should be used for best detection performance.

For the performance comparison among different object detection schemes for the five scenes included for evaluation, Mask R-CNN, Faster R-CNN and YOLO seem to perform equally well except SSD, with each achieving best results (marked in red in Table 2) either in two scenes, or for one scene plus for all five scenes together. As for more detailed comparison among different scenes, only results of (v) are considered for simplicity, which are obtained using image data from all channels according to the fusion scheme described in Sec. II-D. First of all, since both Doorway and Staircase are obtained from indoor scenes with few environmental interferences, good detection results, i.e., AP > 85%, are achieved with all methods. As for the other indoor scenes, i.e., Room and Bus, since some low illumination conditions can be found in both scenes while noticeable ambient lights can also be found in the latter, low APs (<50%) are obtained for the latter by SSD (and some very low APs are found in Channel 2 for the former by YOLO and Faster R-CNN). As for the case of Pathway, relatively low AP values are obtained for the corresponding outdoor scenes because of large variation in object distance, and (v) gives the best results for all models. Such results also can be observed for the All case.

C. FACE TRACKING EVALUATION

Due to ambient light in the IR video, many false positive may occur in face detection. The approach proposed

| Channel | SSD | YOLO | Faster | Mask |
|---------|-----|------|--------|------|
| (i)     | 52.23 | 67.77 | 62.30  | 58.82 |
| (ii)    | 48.44 | 71.02 | 58.84  | 62.78 |
| (iii)   | 54.94 | 82.85 | 65.67  | 69.16 |
| (iv)    | 52.48 | 81.48 | 67.15  | 64.85 |
| (v)     | 59.17 | 84.38 | 69.76  | 74.68 |

TABLE 2. Comparison of AP for different models on different scenes. (i), (ii), (iii) represent Channel 2, 4, 6 respectively. (iv) is the fusion of Channel 2 and 4. (v) is the fusion of all three channels.

5 Nonetheless, (iv) does improve the performance over worst-case single channel results in all 24 scene-model cases and with extended coverage of effective ranges, except for the five marked with underlines.
TABLE 3. Comparison of F-measure scores of face detection and tracking.

| F-measure          | Pathway 1 | Pathway 2-3 |
|--------------------|-----------|-------------|
|                    | ch2       | ch4 | ch6 | ch2 | ch4 | ch6 |
| Detection          | 57.33     | 51.80 | 61.58 | 39.30 | 45.27 | 36.44 |
| +Tracking          | 76.5      | 90.76 | 84.26 | 43.41 | 49.56 | 47.06 |
| +Inter-channel fusion | 91.95     | 55.23 |       |       |       |       |

FIGURE 10. Sample images from Bus, wherein one, two, and three persons are detected in Channels 2, 4, and 6, respectively.

FIGURE 11. Sample images from Staircase, wherein one, two, and two persons are detected in Channels 2, 4, and 6, respectively.

FIGURE 12. Sample images from Room, wherein zero, one, and three persons are detected in Channels 2, 4, and 6, respectively.

FIGURE 13. Sample images from Channel 4 of Pathway, with face tracking results of person 4.

IV. CONCLUSION

This work evaluates state-of-the-art human and face detectors and reports their performances on an existing multi-intensity IR illumination dataset, with complete annotations also established for the dataset. To that end, a baseline approach is proposed, which is based on pre-trained CNN detectors, a recently proposed tracker, and simple fusion scheme to take advantage of the complementary effect among different illumination intensities. While satisfactory detection and tracking results are demonstrated in this paper for some results exhibit improvements in F-measure after adopting the tracking method. For both video clips, the channel with non-extreme intensity of illumination (Channel 4) seems to provide more useful image features in the face and generate best face tracking results. Note that these improvements are much more significant for Pathway 1, which may due to the interaction between the two complex situations (i) and (ii) mentioned at the end of Sec. II-B) associated with Pathway2-3 and need further investigation. Additional results (last row of Table 3) also demonstrate that the fusion scheme proposed in Sec. II-D can further improve the performance of face detections.
simple scenes, further improvements for more complicated datasets, better fusion methods, as well as a systematic way of determining relevant parameters, such as batch size or learning rate for training a specific CNN model, are currently under investigation.

APPENDIX

SAMPLE IMAGES FROM SELECTED DATASETS

See Figures 10–13.

REFERENCES

[1] W. Teng, “A new design of ir illuminator for nighttime surveillance,” M.S. thesis, Dept. Comput. Sci., Nat. Chiao Tung Univ., Hsinchu, Taiwan, 2010.

[2] C.-H. Han, H.-T. Chen, W.-C. Teng, C.-W. Liu, and J.-H. Chuang, “M3: Multi-intensity infrared illumination video database,” in Proc. Vis. Comput. Image Process. (VCIP), Dec. 2015, pp. 1–4.

[3] P.-I. Lu, J.-H. Chuang, and H.-H. Lin, “Intelligent nighttime video surveillance using multi-intensity infrared illuminator,” in Proc. World Congr. Eng. Comput. Sci., vol. 1, 2011, pp. 19–21.

[4] Y.-T. Chen, J.-H. Chuang, W.-C. Teng, H.-H. Lin, and H.-T. Chen, “Robust license plate detection in nighttime scenes using multiple intensity IR-illuminator,” in Proc. IEEE Int. Symp. Ind. Electron., May 2012, pp. 893–898.

[5] J.-H. Chuang, W.-J. Tsai, C.-H. Chan, W.-C. Teng, and I.-C. Lu, “A novel video summarization method for multi-intensity illuminated infrared videos,” in Proc. IEEE Int. Conf. Multimedia Expo. (ICME), Jul. 2013, pp. 1–6.

[6] 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.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961–2969.

[8] 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.

[9] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” 2020, arXiv:2004.10934.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 21–37.

[12] K. Guo, S. Wu, and Y. Xu, “Face recognition using both visible light image and near-infrared image and a deep network,” CAAI Trans. Intell. Technol., vol. 2, no. 1, pp. 39–47, 2017.

[13] S. Cho, N. Baek, M. Kim, J. Koo, J. Kim, and K. Park, “Face detection in nighttime images using visible-light camera sensors with two-step faster region-based convolutional neural network,” Sensors, vol. 18, no. 9, p. 2995, Sep. 2018.

[14] H. Jiang and E. Learned-Miller, “Face detection with the faster R-CNN,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2017, pp. 650–657.

[15] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4293–4302.

[16] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.

[17] Y.-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 Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 740–755.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4510–4520.

[20] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5525–5533.

[21] L. Bourdev, S. Maji, T. Brox, and J. Malik, “Detecting people using mutually consistent poselet activations,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2010, pp. 168–181.

[22] A. RoyChowdhury, P. Chakraborty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller, “Automatic adaptation of object detectors to new domains using self-training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 780–790.

[23] S. Jin, H. Su, C. Stauffer, and E. Learned-Miller, “End-to-end face detection and cast grouping in movies using Erdös–Rényi clustering,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5276–5285.

[24] L. Liu and M. T. Özsu, Encyclopedia of Database Systems, vol. 6. New York, NY, USA: Springer, 2009.

PEGGY JOY LU received the B.S. and M.S. degrees from the Computer Science and Information Engineering Department, National Chiao Tung University, in 2009 and 2011, respectively. She is currently pursuing the Ph.D. degree with the Computer Science and Information Engineering Department, National Yang Ming Chiao Tung University. She is currently working with the National Center for High-Performance Computing. Her research interests include computer vision, federated learning, and blockchain.

JEN-HUI CHUANG (Senior Member, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, in 1980, the M.S. degree in electrical and computer engineering from the University of California at Santa Barbara, in 1883, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Urbana–Champaign, in 1991. Since 1991, he has been on the faculty with the Department of Computer Science, National Chiao Tung University (NCTU). From 2004 to 2005, he was the Chairperson of the Department of Computer and Information Science, NCTU; where he was the Dean of the College of Computer Science, from 2017 to 2020. He is currently the Vice-President of the Academic Affairs with National Yang Ming Chiao Tung University. His research interests include signal and image processing, computer vision and pattern recognition, and robotics. He was the President of the Chinese Image Processing and Pattern Recognition Society, from 2015 to 2016. He served as the Governing Board Member for the International Association of Pattern Recognition, from 2013 to 2016, and the Governing Board Member for the IEEE Taipei Chapter, from 2017 to 2020. He served as an Associate Editor for Signal Processing, from 2005 to 2010. He has been an Associate Editor of Journal of Information and Science and Engineering, since 2009.

* * *