Multi-domain Object Detection Framework using Feature Domain Knowledge Distillation

Da-Wei Jaw, Shih-Chia Huang, *Senior Member, IEEE*, Zhi-Hui Lu, *Member, IEEE*, Benjamin C. M. Fung, *Senior Member, IEEE*, Sy-Yen Kuo, *Life Fellow, IEEE*

***Abstract*—Object detection techniques have been widely studied, utilized in various works, and have exhibited robust performance on images with sufficient luminance. However, these approaches typically struggle to extract valuable features from low-luminance images, which often exhibit blurriness and dim appearence, leading to detection failures. To overcome this issue, we introduce an innovative unsupervised feature domain knowledge distillation framework. The proposed framework enhances the generalization capability of neural networks across both low- and high-luminance domains without incurring additional computational costs during testing. This improvement is made possible through the integration of generative adversarial networks and our proposed unsupervised knowledge distillation process. Furthermore, we introduce a region-based multiscale discriminator designed to discern feature domain discrepancies at the object level rather than from the global context. This bolsters the joint learning process of object detection and feature domain distillation tasks. Both qualitative and quantitative assessments shown that the proposed method, empowered by the region-based multiscale discriminator and the unsupervised feature domain distillation process, can effectively extract beneficial features from low-luminance images, outperforming other state-of-the-art approaches in both low- and sufficient-luminance domains.**

***Index Terms*—Generative Adversarial Networks, Object Detection, Unsupervised Knowledge Distillation.**

I. Introduction

LOCALIZING and identifying objects within images is crucial for a wide range of vision-based applications. Over the past decade, object detection algorithms have undergone extensive research and have been exploited across varied applications, such as human-machine interaction [1], [2], target tracking [3], [4], and surveillance systems [5]–[8].

Modern detection systems localize objects by adopting regression techniques, and concurrently classify them using various categorization approaches. The R-CNN series [9]–[12] and Feature Pyramid Networks (FPNs) [13] employ a two-stage strategy that initially estimates the Region of Interest (RoI) and subsequently classifies the object to achieve object detection. The YOLO series [14]–[16], SSD [17], and RetinaNet [18] have received considerable attention due to their unified approach that performs object classification and localization in parallel, thereby providing satisfactory performance and impressive computational efficiency.

(a) Precision: 67%, Recall: 40% (b) Precision: 60%, Recall: 60%

(c) Precision: 75%, Recall: 60% (d) Precision: 83%, Recall: 100%

Fig. 1. Advance object detection techniques such as (a) RetinaNet [18], (b) YOLOv3 [16], and (c) SFT-Net [19], exhibit demonstrable limitations in accurately identifying the 'Person' class under low luminance conditions. In contrast, our proposed approach (d) showcases a superior capacity for accurate detection under these challenging light conditions.

However, while these methodologies yield satisfactory results on images with sufficient luminance, their performance degrades considerably under low-luminance conditions, such as those encountered during nighttime. Fig. 1 illustrates this limitation. Despite training advanced detectors, like YOLO-v3 [16], RetinaNet [18], and SFT-Net [19], on datasets encompassing both low- and adequate-luminance images, these techniques fail to detect the $Person$ class within low-luminance images.

Such detection failures primarily stem from these algorithms' inability to extract valuable features from low-luminance, often blurred images, leading to ambiguous and indistinct features. One potential solution could be to apply image enhancement techniques prior to the object detection process to improve the visibility of the input image, thereby enhancing the performance of the detection task. However, recent studies [20] have suggested that low-level image processing techniques aimed at enhancing image visibility may

D.-W. Jaw and S.-Y. Kuo is with Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan. E-mail: jdw.davidjaw@gmail.com.

S.-C. Huang is with Department of Electronic Engineering, National Taipei University of Technology, Taipei 106, Taiwan. E-mail: schuang@ntut.edu.tw. (Corresponding author: S.-C. Huang)

Benjamin C. M. Fung is with School of Information Studies, McGill University, Montreal, Canada

Z.-H. Lu is with School of Computer Science, Fudan University, Shanghai 200433, China and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China. Email: lzh@fudan.edu.cn. (Corresponding author: Z.-H. Lu)

This work was supported by the Ministry of Science and Technology of Taiwan under Grant 108-2221-E-027-047-MY3, 110-2221-E-027-046-MY3, 111-2221-E-002-136-MY3, 111-2221-E-002-133-MY3, 111-2811-E-027-008-MY2, 111-2221-E-027-138, 111-2314-B-002-297, 111-2218-E-167-001 and Shanghai Science and Technology Project under Grant (No.22510761000).

not necessarily improve the outcomes of subsequent high-level vision tasks due to the domain discrepancy issue, a perspective also endorsed by our experimental findings.

To address these challenges, we propose a region-based multiscale feature domain distillation framework, referred to as RMD-Net, which enhances the overall generalization ability of the network without imposing additional computational burdens during the testing phase. Specifically, the proposed RMD-Net enables neural networks to glean feature domain knowledge from the beneficial high-luminance feature space. This strategy draws inspiration from knowledge distillation (KD) tasks [21], where a compact model learns from a more complex, deeper network, thereby improving its generalization ability and ultimately enhancing overall performance.

In conventional knowledge distillation processes [21], the student network is trained to minimize the divergence between its predictions and the soft targets produced by the teacher network using identical input data. This technique promotes the learning of feature representations akin to those of the teacher network, thereby improving generalization to unseen data. However, obtaining daytime and nighttime images with identical content is unfeasible in real-world settings.

To circumvent this obstacle, we introduce RMD-Net, which implements an unsupervised knowledge distillation-like mechanism: a teacher network is trained exclusively on images from the high-luminance domain, thereby accumulating pristine, bright feature knowledge. By designating the feature space derived from the teacher network as the target domain, we leverage an unsupervised learning mechanism inspired by the Generative Adversarial Network (GAN) framework. This permits the unsupervised distillation of feature domain knowledge to the student network, leading to enhanced feature extraction across both low- and high-luminance domains. Additionally, we propose a region-of-interest (RoI)-based multiscale network architecture that allows the neural network to focus on the object, as opposed to global or background information.

Employing the proposed training strategy and the tailored region-based multiscale network, our framework is capable of distilling knowledge from the pristine feature domain to the generator network, thereby boosting the network's generalization capacity while retaining the same computational complexity during the testing phase.

This work presents several substantial contributions to the field of object detection in low-luminance scenarios, as summarized below:

- We propose a novel region-based unsupervised feature domain distillation framework, RMD-Net, which melds the principles of knowledge distillation and the Generative Adversarial Network (GAN) framework. RMD-Net allows the knowledge from a teacher network trained on the sufficient-luminance domain to be distilled to the student network that is designed to perform in the challenging low-luminance domain. This novel approach enhances the generalization ability of the student network across both the low-luminance and sufficient-luminance domains. The unsupervised nature of our distillation process opens up new possibilities in cross-domain knowledge transfer.
- We introduce a region-of-interest (RoI)-based multiscale network architecture, which allows the discriminator network to concentrate primarily on the object of interest rather than the global image. This strategy further enhances the object detection performance under low-luminance conditions.
- Our framework significantly surpasses the performance of both the combination of image enhancement and object detection techniques, as well as other state-of-the-art object detection methods. This demonstrates its effectiveness in challenging low-luminance environments. Despite the improvements in performance and generalization ability, our method does not increase the computational complexity during the testing phase. This makes our approach suitable for real-time applications.

In summary, the proposed framework brings an innovative perspective to the task of object detection in low-luminance scenarios, significantly enhancing the generalization ability and performance of the object detection network without introducing additional computational complexity.

The rest of this study is summarized as follows: Section II briefly introduces the object detection approaches and the Generalization Adversarial Networks, Section III details the proposed RMD-Net, Section IV demonstrates the qualitative and quantitative evaluations of the proposed method. Section V concludes the paper.

II. Related Works

A. Object Detection Methods

Existing object detection techniques can be primarily categorized into two categories [18]. The first encompasses those employing a stage-wise approach, initially identifying potential object regions and subsequently classifying the objects during the second stage. The second category simultaneously accomplishes localization and classification within a single stage, yielding satictory accuracy and relatively swift computational efficiency.

1) Two-Stage Object Detection Methods: Within the scope of two-stage methods, the technique proposed by Girshick *et al*. [9] was pioneering in employing selective search [22] for generating region proposals, utilizing CNNs for feature extraction, and adopting a support vector machine (SVM) for object classification. To improve upon R-CNN [9], Fast R-CNN [10] was introduced, utilizing region of interest (RoI) pooling to process all region proposals concurrently rather than individually. Furthermore, SVM was replaced by CNNs to enhance performance. In Faster R-CNN [11], a region proposal network was incorporated to replace the selective search method, creating potential object locations within a learnable mechanism and thereby increasing performance while maintaining impressive computational speed. Feature Pyramid Networks [13] leverage a top-down network hierarchy with lateral connections across various resolutions, leading to significant enhancements in accuracy and computational efficiency, and gaining substantial attention.

2) One-Stage Object Detection Methods: Within the domain of one-stage object detection, YOLO [14] proposed a regression-based approach, enabling real-time object detection. SSD [17] adopted a multiscale prediction strategy to improve accuracy across various object sizes. YOLO-v2 [15] extended its initial version by optimizing network designs through the application of batch normalization [23], high-resolution classifiers, anchor boxes, and direct location prediction methodology. RetinaNet [18] utilized FPN [13] as the foundational network architecture and proposed a loss function, known as *Focal Loss*, to counter the class imbalance issue. YOLO-v3 [16] designed a lightweight network architecture that employs top-down cross-resolution feature fusion, multiscale prediction, and *Focal Loss*, delivering a significant improvement in computational efficiency while maintaining acceptable performance levels.

With respect to one-stage multi-domain object detection methods, Huang $et\ al.$ [24] introduced an object detection framework that enhances object detection accuracy under rainy weather conditions. Their work proposes a feature selection network that learns to select features generated by a rainy image restoration network. Subsequently, a feature absorption network is employed to enable the object detection backbone to learn from the selected features, thereby enhancing its feature interpretability for rainy images. SFT-Net [19] introduces a self-adaptive feature transformation mechanism to improve the performance of low-light object detection by incorporating an additional feature transformation (FT) module and a gating design. This approach allows the network to choose between utilizing the transformed features produced by the FT module or the original features through the gating mechanism, ultimately leading to improved performance.

B. Knowledge Distillation

Knowledge distillation (KD) aims to boost the performance of a compact, smaller model by mimicking the behavior of a larger, more complex network. The seminal work in this domain was pioneered by Hinton $et\ al.$ [21], who successfully enhanced the generalization capacity and performance of the student network using soft labels provided by the teacher network, with identical input images. Subsequent to [21], various distillation works have been proposed from diverse perspectives, such as supervising high-level features via hint learning [25], mimicking the features outlined by an attention map [26], measuring cross-sample similarity [27], and formalizing the KD task as a distribution matching problem [28].

C. Generative Adversarial Networks

Generative adversarial networks (GAN), introduced by Goodfellow $et\ al.$ [29], with the objective of training a generative model that translates a latent space into a desired distribution. Typically, GAN comprises a generator and a discriminator; the former seeks to produce a distribution that is indistinguishable from the target distribution, whereas the latter is trained to classify whether a sample originates from the real distribution or the generated one. The value function is expressed as

$$\begin{aligned}\min_G \max_D (D,G) = \ & \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \\ & \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))],\end{aligned} \tag{1}$$

where D and G denote the discriminator and generator networks, respectively; p_r and p_z denote the real distribution and the latent space, respectively; x denotes the input images sampled from p_r; and z denotes a random vector sampled from latent space p_z.

GANs are notorious for their instability during training. To stabilize the training process, various methodologies have been proposed. Notably, the WGAN series [30], [31] proposed the usage of Wasserstein distance to measure the distance between the generated and target distribution. This approach has attracted substantial attention due to its superior results in image generation tasks and its contribution to stability during training. The loss function defined in WGAN-GP [31] is as follows:

$$\mathcal{L}_D = \mathbb{E}_{z\sim p_z}[D(G(z))] - \mathbb{E}_{x\sim p_r}[D(x)] + \lambda_{GP}\mathcal{L}_{GP}, \tag{2}$$

where the $\mathcal{L}_{GP}$ is the gradient penalty loss to ensure the discriminator network can measure the Wasserstein distance between two distributions:

$$\mathcal{L}_{GP} = \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2], \tag{3}$$

where $\hat{x}$ denotes the linear interpolation between the real sample x and generated sample $G(z)$.

D. Discussion

While the aforementioned object detection techniques yield satisfactory results for images with sufficient luminance, they face challenges in handling the blurry and dim nature of low-luminance images, leading to reduced performance in this domain. For multi-domain object detection methods, Huang's method [24] can detect objects in both normal and rainy domains. However, its framework requires a well-trained domain-to-domain image restoration process to train the object detection network, which is not available in our case. Despite the impressive results achieved by SFT-Net [19], it presents several drawbacks: increased computational complexity, a larger number of parameters for the additional modules required for feature transformation, and a feature transformation mechanism that primarily focuses on the global image, which may ultimately limit its performance.

In this study, our aim is to address the challenging task of low-luminance object detection by developing a region-based, unsupervised feature domain distillation framework, and a training mechanism that boosts the generalization capability of the object detection network without adding any computational complexity during testing. We adopt YOLO-v3 [16] as our backbone architecture, given its computational efficiency and relatively high performance. While KD methods can distill knowledge to the student network with a superior neural network model as its teacher, applying such methodology directly to our multi-domain dataset is infeasible. This is because, while the advantageous feature space has

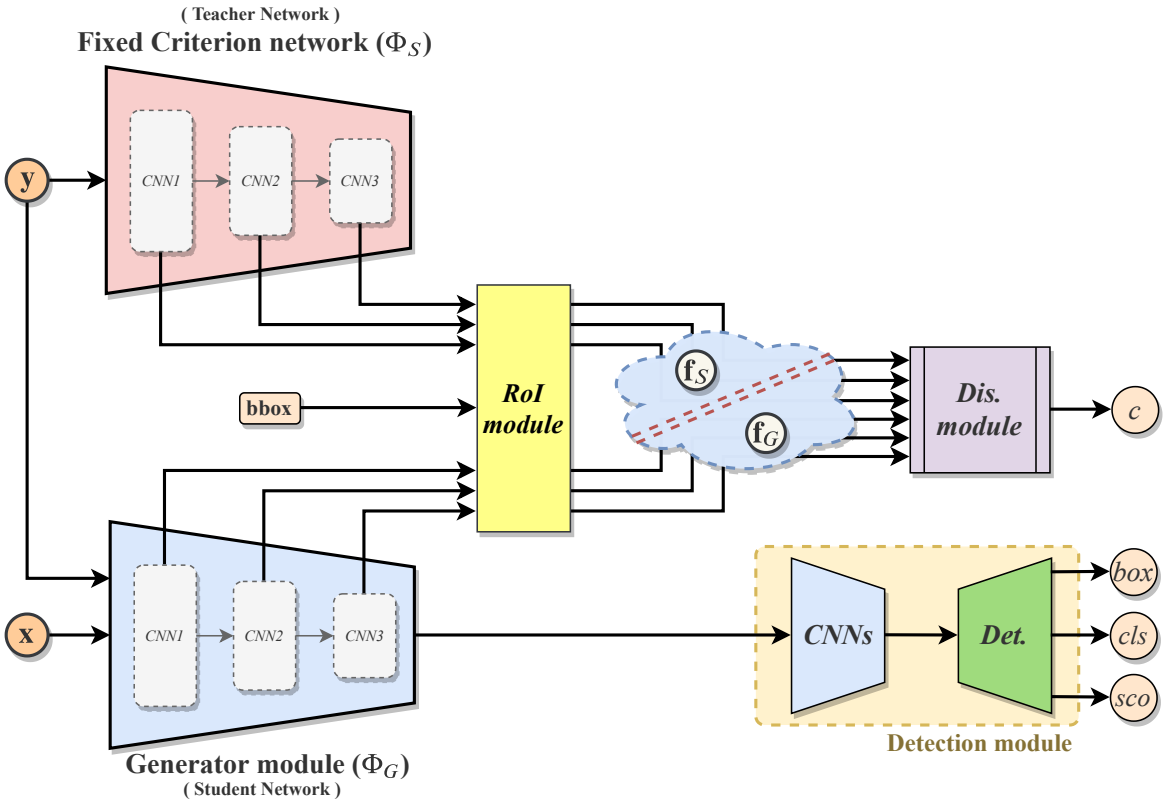

Fig. 2. An overview of the proposed Region-based Multiscale Discriminator Network (RMD-Net) is provided: Here, $\mathbf{x}$ corresponds to images captured under nighttime conditions, $\mathbf{y}$ represents images captured under daytime conditions, **bbox** stands for the ground truth bounding box, while box, cls, sco denote the predictions made by the Detection module. The critics c are estimated by the Discriminator module.

been delineated by training the teacher network exclusively with sufficient-luminance images, the teacher network cannot impart beneficial knowledge when the input images are under low-luminance conditions. Therefore, identical input-based KD methodologies are not applicable to our framework. As a result, we propose the development of an unsupervised distillation framework using the principles of GAN.

Typically, the GAN framework is employed to train a generator network that maps one distribution to a specific image distribution [29]–[31]. However, our objective is not image generation or image domain transformation, but rather the distillation of beneficial knowledge from the sufficient-luminance feature space to the low-luminance feature space. Thus, the GAN framework isn't readily applicable to our framework. To overcome this challenge, we propose an unsupervised domain distillation process that combines the principles of knowledge distillation and the GAN framework, together with a region-based, multiscale discriminator network architecture. This proposed framework not only enhances the generalization capability of neural networks across both domains, but also yields significantly better performance in the task of low-luminance object detection, all while maintaining the same computational complexity during testing.

III. Proposed Method

As depicted in Fig. 2, the proposed RMD-Net consists of four modules: (1) the Generator module and Criterion network, which correspond to the student and teacher networks in typical KD tasks; (2) the Detection module, which classifies and localizes objects on the basis of the features provided by the Generator module; (3) the RoI module, which crops and resizes objects from the provided feature maps; and (4) the Discriminator module, which learns to distinguish features from the target and source domains, thereby providing gradients for the Generator module to learn distillation.

In this work, the images captured during daytime and nighttime conditions are defined as the sufficient-luminance domain ($\mathbb{P}_S$) and low-luminance domain ($\mathbb{P}_L$), respectively. Our purpose is to improve the generalization ability of neural networks in the multi-domain dataset by distilling the knowledge from the sufficient-luminance domain $\mathbb{P}_S$, which has clear and bright features.

To achieve this objective, we first train a Criterion network with only the sufficient-luminance domain ($\mathbb{P}_S$) on the object detection task, thereby obtaining a Criterion network that can produce high-quality features when the input images are from $\mathbb{P}_S$. The Criterion network is denoted as Φ_S. Second, a Generator network with identical structures (denoted as Φ_G) is trained using images from both sufficient- and low- luminance domains ($\mathbb{P}_S$,$\mathbb{P}_L$).

After obtaining Φ_S and Φ_G, we denote the features extracted by each network as $\mathbf{f}_S$ and $\mathbf{f}_G$, respectively. Therefore, an unsupervised learning procedure is employed to distill the knowledge from the feature domain $\mathbb{P}_{\mathbf{f}_S}$ to the feature

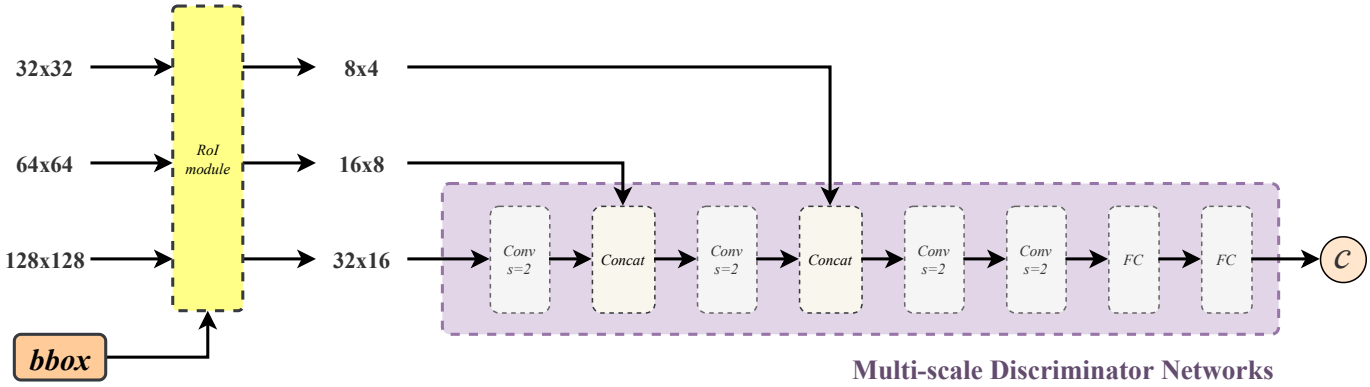

Fig. 3. A detailed illustration of the proposed region-based multiscale discriminator: after object cropping using the RoI module, the discriminator networks fuse features across different resolutions to compute the Wasserstein distance between the target and the projected domain.

domain $\mathbb{P}_{\mathbf{f}_G}$. Specifically, our aim is to ensure that the features produced by our Generator network Φ_G, either estimated from daytime image $\mathbf{x}$ or nighttime image $\mathbf{y}$, to be indistinguishable by the unsupervisedly trained discriminator. In this manner, the quality of feature space $\mathbb{P}_{\mathbf{f}_G}$ can be improved via the feature domain projection process, thus improving the overall performance of our object detection method under low-luminance conditions. The loss function of our feature domain projection procedure is defined as follows:

$$\mathcal{L}_{proj.} = \mathbb{E}_{\mathbf{z}\sim\{\mathbb{P}_S,\mathbb{P}_L\}}[-\Phi_D(\Phi_G(\mathbf{z}))], \tag{4}$$

where $\mathbf{z}$ denotes an image randomly sampled from two domains ($\{\mathbb{P}_S, \mathbb{P}_L\}$), Φ_G denotes the Generator module, and Φ_D denotes the Discriminator module. Notably, in contrast to the traditional GAN framework that takes images as input of the discriminator, we directly take features produced by the Generator module as input to accomplish the feature domain projection.

Correspondingly, to endow the ability of measuring the differences between the source feature domain $\mathbb{P}_{\mathbf{f}_G}$ and targeting feature domain $\mathbb{P}_{\mathbf{f}_S}$, we defined the loss function of our Discriminator module as:

$$\begin{aligned}\mathcal{L}_{Dis.} = \ & \mathbb{E}_{z\sim\{\mathbb{P}_S,\mathbb{P}_L\}}[\Phi_D(\Phi_G(z))] - \mathbb{E}_{y\sim\mathbb{P}_S}[\Phi_D(\Phi_S(y))] \\ & + \lambda_{GP}(\|\nabla_{\tilde{\mathbf{f}}}\Phi_D(\tilde{\mathbf{f}})\|_2 - 1)^2,\end{aligned} \tag{5}$$

where Φ_D, Φ_S denote the Discriminator module and the Criterion network, respectively; λ_{GP} denotes the weighting of the gradient penalty term; and $\tilde{\mathbf{f}}$ denotes the linear interpolation between $\mathbf{f}_S$ and $\mathbf{f}_G$, as defined in the WGAN-GP [31] framework.

A. Training strategy

In order to preserve the information for the object detection task during the training of domain projection, we designed a simple training policy that interactively trains the two tasks, which are depicted in Algorithm 1. In Step 1, Φ_S with images from the sufficient-luminance domain ($\mathbb{P}_S$), and Φ_G with images from two domains ($\mathbb{P}_S, \mathbb{P}_L$) are trained on the object detection task. In Step 2, the Discriminator module with $\mathcal{L}_{Dis.}$ is trained for learning the feature domain discrepancy. In Step 3, the Generator module with $\mathcal{L}_{proj.}$ is trained for the domain projection task. In step 4, the object detection task with images from both domains is trained. If the Generator module does not converge on the basis of the early stopping policy, then Steps 2 to 4 are repeated until the module converges.

Algorithm 1 The proposed training policy. Notably, the object detection loss $\mathcal{L}_{obj.}$ is as defined in the YOLO-v3 [16]. Notably, the S1, S2, S3, S4 refers to training steps 1 to 4 as defined in Section III-A.

Require: Initialize the parameters for Criterion network w_S, Generator module w_G, Discriminator w_D.

while Φ_S has not converged on $\mathcal{L}_{obj.}$ **do** ▷ S1
  Sample sufficient-luminance image $y \sim \mathbb{P}_S$.
  $w_S \leftarrow$Adam(w_S, $\mathcal{L}_{obj.}(\Phi_S,\ y)$).
end while
while Φ_G has not converged on $\mathcal{L}_{obj.}$ **do**
  Sample random image $z \sim \{\mathbb{P}_S, \mathbb{P}_L\}$.
  $w_G \leftarrow$Adam(w_G, $\mathcal{L}_{obj.}(\Phi_G,\ z)$).
end while
while Φ_G has not converged on $\mathcal{L}_{obj.}$ **do**
  for $i = 1, \ldots, n_{critic}$ **do** ▷ S2
    Sample sufficient-luminance image $y \sim \mathbb{P}_S$, random image $z \sim \{\mathbb{P}_S, \mathbb{P}_L\}$.
    $\mathbf{f}_S \leftarrow \Phi_S(y)$.
    $\mathbf{f}_G \leftarrow \Phi_G(z)$.
    $w_D \leftarrow$Adam($w_D, \mathcal{L}_{Dis.}(\Phi_D, \mathbf{f}_S, \mathbf{f}_G)$)
  end for
  Sample random image $z \sim \{\mathbb{P}_S, \mathbb{P}_L\}$. ▷ S3
  $\mathbf{f}_G \leftarrow \Phi_G(z)$.
  $w_G \leftarrow$Adam($w_G, \mathcal{L}_{proj.}(\Phi_D, \mathbf{f}_G)$).
  $w_G \leftarrow$Adam($w_G, \mathcal{L}_{obj.}(\Phi_G, \mathbf{f}_G)$). ▷ S4
end while

B. Region-based multiscale discriminator

As shown in Fig. 3, the proposed Discriminator module takes feature output from the RoI module as input and fuses features from different resolutions to estimate the domain difference between $\mathbb{P}_S$ and $\mathbb{P}_G$.

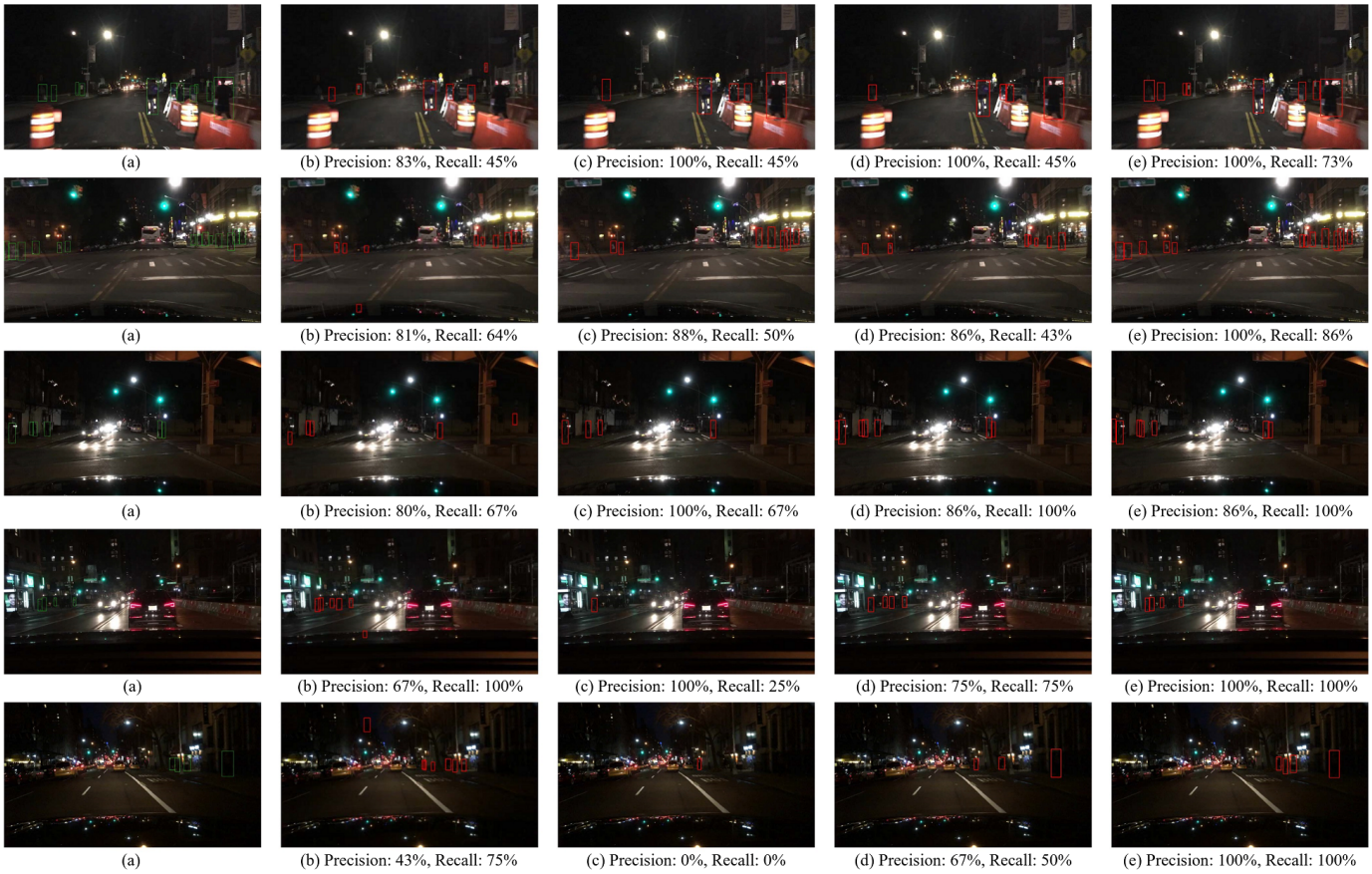

Fig. 4. Comparison between state-of-the-art object detection techniques, including RetinaNet [18], YOLO-v3 [16] and SFT-Net [19]. Our proposed method manifests considerable improvement relative to these techniques.

TABLE I
THE DETAIL NETWORK ARCHITECTURE OF THE PROPOSED MULTISCALE DISCRIMINATOR. IT IS NOTEWORTHY THAT ALL CONV2D LAYERS UTILIZE A FILTER SIZE OF 5.

Type	Filters	Stride	Activation	Dropout	Output Size
Conv2D	128	2	-	-	(16, 8, 128)
Concat	-	-	Leaky ReLU	0.25	(16, 8, 256)
Conv2D	256	2	-	-	(8, 4, 256)
Concat	-	-	-	-	(8, 4, 512)
ZeroPadding	-	-	-	-	(9, 5, 512)
BatchNorm	-	-	Leaky ReLU	0.25	(9, 5, 512)
Conv2D	512	2	-	-	(5, 3, 512)
BatchNorm	-	-	Leaky ReLU	0.25	(5, 3, 512)
Conv2D	512	2	-	-	(3, 2, 512)
BatchNorm	-	-	Leaky ReLU	0.25	(3, 2, 512)
Flatten	-	-	-	-	(3072)
Dense	128	-	-	-	(128)
BatchNorm	-	-	Leaky ReLU	-	(128)
Dense	1	-	-	-	(1)

In particular, the RoI module crops objects from the input feature maps on the basis of the ground truth using the RoI alignment [12] technique; in this manner, the Discriminator module can learn the domain discrepancy directly from the region of the object instead of learning from the global information. Such design not only aligns closely with our purpose, it also provides a stabilized training procedure. As for the target size of cropped objects, we use the average ratio of objects in the dataset as the default value.

Meanwhile, we designed a pyramid hierarchy to fuse features from three different resolutions, thus providing information in various spatial and semantic levels. Features from shallow layers are down-sampled with strided convolutions and then concatenated with features from deep layers to aggregate information, thereby improving the interpretability of the Discriminator module.

IV. EXPERIMENTAL RESULTS

A. Implementation details

1) Generator module: The resolution of each intermediate feature map sent into the Discriminator module is shown in Fig. 3. These feature maps are estimated from the first Residual block (128×128), the third Residual block (64×64), and the 11^{th} Residual block (32×32), respectively, in the YOLO-v3 [16] network.

2) Discriminator module: The overview of the proposed multi-scale discriminator network is shown in Fig. 3, and the detailed network architecture setup is shown in Table I.

3) Others: All the experiments are conducted on a computer with an NVIDIA GTX 1070, an Intel Core i5 CPU 6500, and 16 GB RAM. The compared approaches (SSD [17], RetinaNet [18], YOLO-v3 [16], and SFT-Net [19]) are pretrained on the MS-COCO dataset [32] and fine-tuned on

TABLE II
A COMPARATIVE PERFORMANCE ANALYSIS OF THE PROPOSED METHOD USING A STANDARD DISCRIMINATOR AND THE PROPOSED REGION-BASED MULTISCALE DISCRIMINATOR IN TERMS OF AP. $\mathbb{P}_S$ DENOTES THE SUFFICIENT-LUMINANCE DOMAIN, WHEREAS $\mathbb{P}_L$ REPRESENTS THE LOW-LUMINANCE DOMAIN.

Model name	Ablation type		AP		
	Multiscale feature	RoI module	$\mathbb{P}_S$	$\mathbb{P}_L$	overall
YOLO-v3 [16]			59.51	41.62	50.14
baseline			61.47	47.32	54.06
RMD-MS	✓		62.88	50.18	56.23
RMD-Net	✓	✓	63.74	53.12	58.18

TABLE III
A COMPARATIVE PERFORMANCE ANALYSIS OF STATE-OF-THE-ART OBJECT DETECTION TECHNIQUES IN TERMS OF AP. $\mathbb{P}_S$ CORRESPONDS TO THE SUFFICIENT-LUMINANCE DOMAIN, WHILE $\mathbb{P}_L$ REPRESENTS THE LOW-LUMINANCE DOMAIN.

Method	$\mathbb{P}_S$	$\mathbb{P}_L$	Overall
SSD-512 [17]	55.48	35.53	45.04
RetinaNet [38]	62.91	33.56	47.54
YOLO-v3 [16]	59.51	41.62	50.14
SFT-Net [19]	61.83	47.43	54.29
RMD-Net	**63.74**	**53.12**	**58.18**

TABLE IV
A COMPARATIVE ANALYSIS BETWEEN THE COMBINED APPROACH OF OBJECT DETECTION AND IMAGE ENHANCEMENT, IN TERMS OF AP. $\mathbb{P}_S$ DENOTES THE SUFFICIENT-LUMINANCE DOMAIN, WHEREAS $\mathbb{P}_L$ REPRESENTS THE LOW-LUMINANCE DOMAIN.

Method	$\mathbb{P}_S$	$\mathbb{P}_L$	Overall
YOLO-v3 [16]	59.51	41.62	50.14
YOLO-v3 + DSLR [34]	59.51	37.83	50.14
YOLO-v3 + SID [35]	59.51	30.03	44.08
RMD-Net	**63.74**	**53.12**	**58.18**

our dataset using the RMSProp optimizer [33] with a learning rate of $5e^{-5}$. As for the image enhancement techniques in our experiments (DSLR [34] and SID [35]), we use the well-trained module.

B. Dataset

In this study, we use the multi-domain dataset proposed by SFT-Net [19], which combines data from the KITTI [36] and BDD-100K [37] datasets, as well as images captured by the authors themselves. This dataset contains images captured during the daytime or nighttime, which aligns with our objective of distilling knowledge from the daytime feature domain to the nighttime domain. Additionally, the authors re-labeled all nighttime images and other images with low-quality labels to ensure high-quality annotations for low-luminance images. In total, the dataset comprises 16,508 images, including 7,865 daytime images (sufficient-luminance domain) and 8,643 nighttime images (low-luminance domain).

C. Comparison

1) Comparison between object detection techniques: To evaluate the proposed method, we compare it with state-of-the-art object detection techniques, including SSD512 [17], RetinaNet [38], YOLO-v3 [16], and the state-of-the-art low-light object detection technique, SFT-Net [19]. Quantitative evaluations are conducted and presented in Table III. The proposed method achieves substantial performance improvements in both domains, demonstrating that the proposed training strategy and region-based multiscale discriminator can successfully transfer high-quality feature knowledge to the generator. Notably, although SFT-Net [19] exhibits enhancements in both sufficient- and low-luminance domains, its performance is still inferior to the proposed method while it incurs additional computational complexity during the testing phase.

The proposed RMD-Net maintains an identical network architecture compared to YOLO-v3 [16] in the test phase, yet our performance significantly improves by over 4.25 AP and 11.50 AP in the sufficient-luminance domain ($\mathbb{P}_S$) and the low-luminance domain ($\mathbb{P}_L$), respectively. The qualitative evaluation of these methods is illustrated in Fig. 4, revealing that the proposed method can detect considerably dark objects, while the other methods struggle to process low-luminance features.

2) Comparison between the combination of image enhancement and object detection: This work focuses on addressing the challenge of object detection in low-luminance conditions. In our experiments, we also compared our method with the combination of image enhancement techniques and the YOLO-v3 [16] method. In this experiment, images under low-luminance conditions were enhanced using different image enhancement techniques, including the DSLR [34] and SID [35] methods. The results, presented in Table IV, show that the performance of YOLO-v3 significantly degrades with these image enhancement methods, even more so than the original YOLO-v3 method without any enhancement. This performance degradation could be attributed to the domain discrepancy between enhanced and sufficient-luminance images, an issue also explored in the study by Li et al. [20].

3) Ablation study: We conducted an ablation study to evaluate the effectiveness of our proposed region-based multiscale discriminator, and presented the corresponding results in Table II. As shown, when we removed the multiscale feature and the RoI module design from the discriminator (labelled as $baseline$), the discriminator could only use the least-scale feature to learn domain knowledge. Similarly, the *RMD-MS* design excludes the RoI module from the discriminator. Despite both designs yielding better performance than the YOLO-v3 [16] method, they exhibited significant performance degradation compared to the proposed region-based multiscale

discriminator.

V. Conclusions

This paper has introduced a novel unsupervised learning approach to transfer features from the low-luminance domain to the sufficient-luminance domain. Our method utilizes a Criterion network trained exclusively on the sufficient-luminance domain to establish a beneficial feature space for object detection tasks, which then guides the feature domain knowledge transfer to the generator within a generative adversarial network framework. In contrast to conventional GANs, our framework is tailored to generate feature maps based on distilled knowledge within the designated beneficial feature space. In order to improve the object detection task through effective domain projection learning, we introduced a region of interest (RoI)-based multiscale discriminator. By integrating the RoI module and multiscale network architecture, the discriminator is able to focus on object-specific regions across various resolutions, eschewing the global perspective that often includes irrelevant background details. Our experimental results confirmed the efficacy of our approach. The proposed methodology outperformed both contemporary multi-domain object detection techniques and low-light detection approaches on a multi-domain dataset. It also showed a significant improvement compared to the combination of image enhancement and object detection techniques. To the best of our knowledge, this work represents the first successful application of an unsupervised feature domain knowledge distillation framework to low-luminance object detection.

References

[1] A. Carfì and F. Mastrogiovanni, "Gesture-based human-machine interaction: Taxonomy, problem definition, and analysis," *IEEE Transactions on Cybernetics*, pp. 1–17, 2021.

[2] J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju, "Deep temporal model-based identity-aware hand detection for space human–robot interaction," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 738–13 751, 2022.

[3] S. Javed, A. Mahmood, J. Dias, L. Seneviratne, and N. Werghi, "Hierarchical spatiotemporal graph regularized discriminative correlation filter for visual object tracking," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 12 259–12 274, 2022.

[4] S. Zhang, W. Lu, W. Xing, and L. Zhang, "Learning scale-adaptive tight correlation filter for object tracking," *IEEE Transactions on Cybernetics*, vol. 50, no. 1, pp. 270–283, 2020.

[5] T.-H. Le, S.-C. Huang, and D.-W. Jaw, "Cross-resolution feature fusion for fast hand detection in intelligent homecare systems," *IEEE Sensors Journal*, vol. 19, no. 12, pp. 4696–4704, 2019.

[6] J. Xie, C. Gao, J. Wu, Z. Shi, and J. Chen, "Small low-contrast target detection: Data-driven spatiotemporal feature fusion and implementation," *IEEE transactions on cybernetics*, vol. 52, no. 11, pp. 11 847–11 858, 2021.

[7] C. Huang, Z. Yang, J. Wen, Y. Xu, Q. Jiang, J. Yang, and Y. Wang, "Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 834–13 847, 2021.

[8] L. Cui, P. Lv, X. Jiang, Z. Gao, B. Zhou, L. Zhang, L. Shao, and M. Xu, "Context-aware block net for small object detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2300–2313, 2022.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

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

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[14] 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.

[15] 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.

[16] Redmon, Joseph and Farhadi, Ali, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[19] S.-C. Huang, Q.-V. Hoang, and D.-W. Jaw, "Self-adaptive feature transformation networks for object detection in low luminance images," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 1, pp. 1–11, 2022.

[20] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3838–3847.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, vol. 37, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[24] S.-C. Huang, Q.-V. Hoang, and T.-H. Le, "Sfa-net: A selective features absorption network for object detection in rainy weather conditions," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[25] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6550

[26] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[27] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 2852–2859. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17147

[28] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *CoRR*, vol. abs/1707.01219, 2017. [Online]. Available: http://arxiv.org/abs/1707.01219

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 06–11 Aug 2017, pp. 214–223.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[32] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.
[33] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.
[34] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "Dslr-quality photos on mobile devices with deep convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3277–3285.
[35] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.
[36] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
[37] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

Da-Wei Jaw received the B.S. degree in electronic engineering from National Taipei University of Technology, Taipei, Taiwan, in 2015, and the M.S. degree from the electronic engineering from National Taipei University of Technology, Taipei, Taiwan, in 2017. He is currently working toward the Ph.D. degree in Electrical Engineering at National Taiwan University, with research interests relating to digital image processing, machine learning and neural networks.

Shih-Chia Huang is a Full Professor with the Department of Electronic Engineering at National Taipei University of Technology, Taiwan, and an International Adjunct Professor with the Faculty of Business and Information Technology at the University of Ontario Institute of Technology, Canada. He has been named a senior member of the Institute of Electrical and Electronic Engineers (IEEE). He is currently the Chair of the IEEE Taipei Section Broadcast Technology Society, and was a Review Panel Member of the Small Business Innovation Research (SBIR) program for the Department of Economic Development of Taipei City and New Taipei City, respectively.

Professor Huang has published more than 80 journal and conference papers and holds more than 60 patents in the United States, Europe, Taiwan, and China. Dr. Huang received B.S. and M.S. degrees from National Taiwan Normal University and National Chiao Tung University, respectively. In 2009, Dr. Huang received a doctorate degree in Electrical Engineering from National Taiwan University, Taiwan. He was presented with the Kwoh-Ting Li Young Researcher Award in 2011 by the Taipei Chapter of the Association for Computing Machinery, the 5th National Industrial Innovation Award in 2017 by the Ministry of Economic Affairs, Taiwan, as well as the Dr. Shechtman Young Researcher Award in 2012 by National Taipei University of Technology. Professor Huang was the recipient of an Outstanding Research Award from National Taipei University of Technology in 2014 and the College of Electrical Engineering and Computer Science, National Taipei University of Technology in 2014-2016.

In addition, he has been an associate editor of the Journal of Artificial Intelligence and a guest editor of the Information Systems Frontiers and the International Journal of Web Services Research. He is also the Services and Applications Track Chair of the IEEE CloudCom 2016-2017 conference, the Applications Track Chair of the IEEE BigData Congress in 2015, General Chair of the 2015-2016 IEEE BigData Taipei Satellite Session, and the Deep learning, Ubiquitous and Toy Computing Minitrack Chair of the 2017-2018 Hawaii International Conference on System Sciences.

His research interests include intelligent multimedia systems, image processing and video coding, video surveillance systems, cloud computing and big data analytics, artificial intelligence, and mobile applications and systems.

Zhihui Lu LU is a Professor at School of Computer Science, Fudan University. He received a Ph.D. computer science degree from Fudan University in 2004, and he is a member of the IEEE and China computer federation's service computing specialized committee. His research interests are cloud computing and service computing technology, big data architecture, edge computing, and IoT distributed system.

Benjamin C. M. Fung (Senior Member, IEEE) received the Ph.D. degree in computing science from Simon Fraser University, Canada, in 2007. He is currently the Canada Research Chair of data mining for cybersecurity and a Professor with the School of Information Studies, McGill University, Canada. He has more than 130 refereed publications, with more than 11,000 citations, that span the research forums of data mining, privacy protection, cybersecurity, services computing, and building engineering. He serves as an Associate Editor for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and Sustainable Cities and Society (Elsevier). He is also a licensed Professional Engineer of software engineering in the Province of Ontario, Canada.

Sy-Yen Kuo (Life Fellow, IEEE) received the B.S. degree in electrical engineering from the National Taiwan University (NTU), Taiwan, in 1979, the M.S. degree in electrical and computer engineering from the University of California at Santa Barbara in 1982, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign (UIUC) in 1987. He was a Faculty Member with the Department of Electrical and Computer Engineering, The University of Arizona, from 1988 to 1991, and an Engineer at Fairchild Semiconductor, Sunnyvale, CA, USA, and Silvar-Lisco, Menlo Park, CA, USA, from 1982 to 1984. He was the Chairperson of the Department of Electrical Engineering, NTU, from 2001 to 2004, and the Dean of the College of Electrical Engineering and Computer Science, NTU, from 2012 to 2015. He is currently a Distinguished Professor at the Department of Electrical Engineering, NTU. He has published 450 papers in journals and conferences, and holds 22 U.S. patents, 23 Taiwan patents, and 15 patents from other countries. His current research interests include dependable and secure systems, the Internet of Things, and image processing. He was a member of the IEEE Computer Society Board of Governors from 2017 to 2019. He received the Distinguished Academic Achievement Alumni Award from the Department of Computer Science, University of Illinois Urbana–Champaign, in 2019, and the Distinguished Research Award and the Distinguished Research Fellow Award from the Ministry of Science and Technology in Taiwan. He was also a recipient of the Best Paper Awards from the International Symposium on Software Reliability Engineering in 1996 and the IEEE/ACM Design Automation Conference in 1986, and the U.S. National Science Foundation's Research Initiation Award in 1989. He was the Vice President of the IEEE Computer Society in 2020.