Intelligent Small Object Detection for Digital Twin in Smart Manufacturing With Industrial Cyber-Physical Systems

Xiaokang Zhou, *Member, IEEE*, Xuesong Xu, *Member, IEEE*, Wei Liang, *Member, IEEE*, Zhi Zeng, Shohei Shimizu, Laurence T. Yang, *Fellow, IEEE*, and Qun Jin, *Senior Member, IEEE*

***Abstract*—Recently, along with several technological advancements in cyber-physical systems, the revolution of Industry 4.0 has brought in an emerging concept named digital twin (DT), which shows its potential to break the barrier between the physical and cyber space in smart manufacturing. However, it is still difficult to analyze and estimate the real-time structural and environmental parameters in terms of their dynamic changes in digital twinning, especially when facing detection tasks of multiple small objects from a large-scale scene with complex contexts in modern manufacturing environments. In this article, we focus on a small object detection model for DT, aiming to realize the dynamic synchronization between a physical manufacturing system and its virtual representation. Three significant elements, including equipment, product, and operator, are considered as the basic environmental parameters to represent and estimate the dynamic characteristics and real-time changes in building a generic DT system of smart manufacturing workshop. A hybrid deep neural network model, based on the integration of MobileNetv2, YOLOv4, and Openpose, is constructed to identify the real-time status from physical manufacturing environment to virtual space. A learning algorithm is then developed to realize the efficient multitype small object detection based on the feature integration and fusion from both shallow and deep layers, in order to facilitate the modeling, monitoring, and optimizing of the whole manufacturing process in the DT system. Experiments and evaluations conducted in three different use cases demonstrate the effectiveness and usefulness of our proposed method, which can achieve a higher detection accuracy for DT in smart manufacturing.**

***Index Terms*—Deep neural network, digital twin, industrial cyber-physical systems (CPS), object detection, posture recognition.**

I. Introduction

NOWADAYS, the rapid developments of cyber-physical systems (CPS) and Internet of Things (IoT) in Industry 4.0 [1] have enabled an emerging virtual representation technology, called digital twin (DT), which acts as a bridge to create high connection, integration, and cooperation between the physical and virtual world. Coupled with AI techniques and big data analytics, DT becomes a significant way to realize the rapid analysis and real-time decision-making not only for smart manufacturing [2], [3], but also in modern industrial automation and control systems [4]. More importantly, as the digitalization of machinery and production systems becomes the basis of smart manufacturing, DT is viewed as the biggest technology trend and most promising technological direction for realizing real-time monitoring, diagnosis, prognosis, and maintenance in Industry 4.0 [5].

DT equipped with CPS is characterized as a strong interlinkage between the real world and its digital representation for smart manufacturing [6]. In this paradigm, DT is defined as a digital simulation model that can collect data from the physical space, and trigger actions on the physical equipment simultaneously. In 2003, Michael Grieves first presented the innovative concept of DT on product life-cycle management. Recently, with the development of industrial IoT and AI technologies, DT becomes a new idea of tangible assets in industrial CPS, which results in a typical simulation process by making use of quantifiable model, soft simulation, and production data. Differing from traditional notions of digital modeling and simulation, DT has been identified as not only an efficient way for virtual duplication

Manuscript received August 29, 2020; revised November 12, 2020 and January 24, 2021; accepted February 2, 2021. Date of publication February 25, 2021; date of current version October 27, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 72088101, Grant 91846301, Grant 72091515, and Grant 62072171, in part by the National Key R&D Program of China under Grant 2017YFE0117500, Grant 2019YFE0190500, Grant 2020YFC0832700, and Grant 2019GK1010, and in part by the Natural Science Foundation of Hunan Province of China under Grant 2019JJ40150 and Grant 2018JJ2198. Paper no. TII-20-4121. *(Corresponding authors: Xuesong Xu; Wei Liang.)*

Xiaokang Zhou is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: zhou@biwako.shiga-u.ac.jp).

Xuesong Xu, Wei Liang, and Zhi Zeng are with the Base of International Science and Technology Innovation and Cooperation on Big Data Technology and Management, Hunan University of Technology and Business, Changsha 410205, China (e-mail: xuxs@hutb.edu.cn; weiliang@csu.edu.cn; zhizeng416416@163.com).

Shohei Shimizu is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: shohei-shimizu@biwako.shiga-u.ac.jp).

Laurence T. Yang is with the Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada (e-mail: ltyang@stfx.ca).

Qun Jin is with the Faculty of Human Sciences, Waseda University, Tokorozawa 359-1192, Japan (e-mail: jin@waseda.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3061419.

Digital Object Identifier 10.1109/TII.2021.3061419

of a physical system, but also a key innovation in real-time visualization. Researches have shown evidences [7] that combining DT and CPS in a computational collaboration system could efficiently assist people to better understand the manufacturing process, and add resilience during an evidence-based decision-making process. It is said that DT is playing an increasingly significant role for the next generation of digitalized industry, and will become the building blocks of future smart factories. However, it is still a challenging issue when dealing with the interoperability, dependability, sustainability, reliability, security, and predictability for the cyber-physical integration in smart manufacturing [8], [9].

Previous research works usually built DT models using product, equipment, or service data. However, in addition to sensors, interfaces, controllers, and communications, work staffs and their behaviors should be viewed as significant elements to simulate and evaluate the virtual environment in smart manufacturing. In some high-risk workshops with complex mechanical and electrical production, the high density of personnel may cause safety hazards, and thus calls for strict requirements on the distribution of personnel in different areas. Human workers need to be recognized as an important kind of available manufacturing resources, and even an essential factor in terms of the real-time structural and environmental parameters of a physical asset [10], when building a DT model to predict, estimate, and analyze the dynamic changes for optimizations of the whole manufacturing process.

Accordingly, when human factors are involved in the virtual models to monitor and evaluate dynamic changes of machining conditions and manufacturing resources, it becomes a typical problem of small object detection because the image of workers is relatively small in some large workshops. Due to the different postures, expressions, and illuminations, a robust model is necessary to distinguish the background and staff accurately and quickly. Particularly, DT based on workers' postures and behaviors may focus on the macro and micro levels, respectively. At the macro level, staffs can be viewed as the particle of movement, which ignores the details such as the body, and focuses more on people with the workshop location, distribution, and activity track information. At the micro level, the body movements, including personal postures and behavioral characteristics in relative fixed position such as production stations, are taken into account. Therefore, an efficient fusion mechanism needs to be designed to seamlessly integrate the collected multidimensional sensing data for dynamic evaluations during digital twinning. A smart strategy is necessary to enhance the DT model with deep learning schemes, to learn more precise features during simulations [11], which may achieve better real-time monitoring, controlling, optimization, and rapid prediction with high-level control functions and data exchange modules.

To improve the smart manufacturing based on a better integration of cyber and physical space, we aim to realize a DT-enabled dynamic synchronization for physical objects during manufacturing processes under large-scale scenes. This task requires the real-time recognition for multiple objects with different positions and sizes and then virtually represents them in DT. Conventional machine learning models can hardly tackle this situation, especially for small objects from a large-scale scene with complex contexts. Therefore, a small object detection model for DT (SOD-DT), aiming at capturing the precise environmental features and real-time changes from physical space to virtual space, is proposed to overcome the shortcomings of conventional approaches. Specifically, a hybrid deep neural network is constructed to accurately identify the real-time status of three important targets, namely equipment, product, and operator, as the basic environmental parameters in building a generic DT system of smart manufacturing workshop, which can efficiently support the surveillance of equipment positioning, personnel distribution, and product trajectory based on digital twinning. The main contribution of this article is concluded as follows.

1) A framework of intelligent small object detection for DT is designed, in which the equipment, product, and operator are considered as three basic environmental parameters in DT to analyze and estimate the dynamic characteristics and real-time changes from physical manufacturing space to virtual space.
2) A hybrid neural network model is constructed based on a combination of advantages of MobileNetv2, YOLOv4, and Openpose, in which the depthwise separable convolutions of MobileNetv2 are integrated into YOLOv4 (this part is later referred to as a newly structured network called YOLOv4-M2) and replaces the original CSPDarknet53, to improve the feature extraction and further benefit the static small object detections (e.g., equipment, product), while the Openpose is improved for long-distance human posture recognition based on newly generated feature maps from the integrated YOLOv4-M2, instead of the original VGG-19.
3) An efficient learning algorithm is developed for multitype small object detection based on the feature integration and fusion from both shallow and deep layers, which can be used to model, monitor, and optimize the whole smart manufacturing process in DT system.

The rest of this article is organized as follows. Section II presents a review of the latest literatures related to this article. In Section III, a framework of intelligent small object detection is introduced. The implementation of the proposed model and the object detection algorithm are discussed in Section IV. In Section V, we address the experiment and evaluation results using a real-world dataset. Finally, Section VI concludes this article.

II. Related Work

In this section, several related issues, including DT technology in CPS, and machine learning used in object detections, are reviewed, respectively.

A. DT Technology in CPS

DT is becoming an important technology when mapping physical space to cyber space in DT-enhanced human–machine interface, so as to optimize the decision-making ability of production management or smart manufacturing [12]. The CPS interface can be used for data insertion and data visualization during digital twinning in a data-driven way [13]. Based on DT-enhanced simulations in CPS, environment factors and human behaviors can be effectively regulated during the manufacturing process. In addition, due to the high self-awareness in CPS, human abilities and behaviors have become key factors when modeling the intelligent system based on DT technology. For example, Zhou *et al.* [14] presented a human–cyber-physical system by integrating human behaviors in intelligent manufacturing, which was considered as a new generation of digital manufacturing with three main factors: human, network, and physical system. IoT sensors and cameras were used to design an IoT-based DT for energy-efficient CPS [15], which could improve the work efficiency based on human knowledge management, transfer, and application in smart manufacturing.

Generally, one important step in DT is to create a virtual model to truly reproduce the geometry, attributes, behaviors, and rules of physical entities. Dai *et al.* [16] proposed a DT network which combined the DT with industrial IoT network for the modeling of network topology. They developed a deep reinforcement learning mechanism to deal with the computation offloading and resource allocation problem. Schluse *et al.* [17] focused on experimental DTs, which were used in virtual testbeds for simulations of hybrid application scenarios. Cai *et al.* [18] discussed the development of virtual machine tools based on DTs. They aimed to utilize the manufacturing data and sensory data to model the machine-specific features, which might benefit the diagnosis and prognosis in cyber-physical manufacturing. Leng *et al.* [19] utilized the DT technology to handle the security issue in industrial IoT environments. They built a blockchain-based manufacturing framework, in which a DT model was employed to synchronize the physical and cyber systems. However, recent researches have few considerations on characteristics of multiple objects in terms of their real-time changes during the whole manufacturing process, especially when dealing with the multisource, heterogeneity, large-scale, and high-noise scenes of DT in industrial CPS. Data collected from multiple sensors may need to be efficiently fused, to improve the robustness and reliability, and enhance the model expansibility for DT.

B. Machine Learning for Object Detection

The emergence of machine learning technology, such as CNN, has greatly improved the performance of object detection. For example, Wu *et al.* [20] introduced a so-called funnel-structured cascade detection framework, in which distributed classifiers were built to extract shape-indexed features for multiview face detection. Jang *et al.* [21] built a task-specific architecture to handle the face-related classification based on single-shot learning analysis. They used the fully convolutional neural network with two parallel branches to facilitate detections of multiple objects with different sizes. Building deeper and more efficient learning models is a primary trend to solve detection tasks of multiple objects. Ren *et al.* [22] integrated a region proposal network with Faster R-CNN, to improve the detection of high-quality region based on full-image convolutional features. Zhu *et al.* [23] proposed a multiple classification method based on unsupervised learning. They designed an integrated framework, which could realize the object localization, class discovery, and detector training simultaneously. Tang *et al.* [24] considered visual and semantic similarities into a weakly supervised learning process, which could improve the detection performance based on the highlight of category-specific differences. Shen *et al.* [25] introduced a two-stage leaning model based on the fully convolutional neural network for multitask learning. They used two scale-associated side outputs in each stage of neural network, to improve the efficiency in extracting skeleton pixels based on multiple scales. Han *et al.* [26] presented a specific Bayesian-based framework for geospatial object detection, in which a weakly supervised learning model was constructed to identify the high-level features from spatial and structural information. Sangineto *et al.* [27] focused on the design of a training protocol based on self-paced learning. They considered the reliability between different subsets of data during the training process, and used the fully-supervised Faster R-CNN architecture to build the deep network based classifier for weakly supervised object detection.

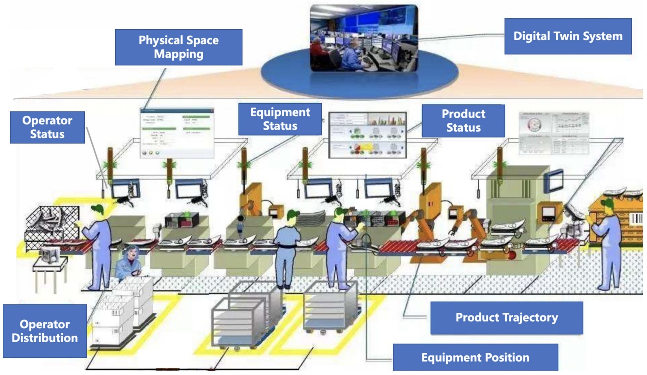


Fig. 1. Digital twinning for smart manufacturing workshop assembly line.

III. Framework of Small Object Detection for DT

In this section, we first introduce two important issues on real-time target recognition in a DT scenario for smart manufacturing workshop. The basic framework for intelligent small object detection in digital twinning is then discussed with two core network modules.

A. Problem Scenario

DT builds a complex system with mutual mapping, timely interaction, and efficient collaboration among human, machine, and environment between physical and virtual space to achieve an on-demand response. It needs to precisely describe the proximity of digital models and physical entities. Thus, three important elements, the equipment, product, and operator, are considered as the basic environmental parameters in terms of their dynamic characteristics and real-time changes, to feed back to the virtual space through various sensors, in order to model, evaluate, and optimize the whole manufacturing process during digital twinning. As shown in Fig. 1, the equipment, workpieces, and operators are included in the digital twinning of a smart workshop floor assembly line. Video or picture sequences captured through surveillance cameras are utilized to provide high-fidelity information to quickly detect and recognize these targets for DT. Specifically, the following two issues on real-time target recognition in practical applications are focused on, to enhance the accuracy of simulation and prediction results in DT system.

1) *Multitype Small Object Detection:* In practical manufacturing scenarios, due to the different spatial dimensions of the site, cameras are usually installed in the farther and higher parts of the workshop, to capture the structural and environmental information of all the targets for DT. Objects in the image are about 10–30 pixels, which can be viewed as a typical detection problem of small targets with multiple types.
2) *Long-Distance Human Posture Recognition:* As one significant dynamic environmental parameter in digital twinning, the operator's behavior is highly autonomous and uncertain. Due to the different angles and distances of cameras, operators' whole-body features are usually unavailable. Traditional algorithms based on key points of human skeleton are prone to be difficult for existing virtual entity models to objectively depict the physical objects.

Summarily, when the target pixel is too small and the feature information is sparse, most of the current convolution operations of deep learning are performed in regions with low target expectations, which often leads to a large waste of computational resources and low execution efficiency. This kind of images of small targets has few features after multilayer convolution processing, which becomes extremely difficult to meet the needs of detection and regression, especially when handling changes in ambient light and smoke from a complex background environment. Therefore, it is necessary to improve the robustness of the detection algorithm, and realize the mapping and interaction between physical and digital space within an acceptable time in digital twinning.

B. Framework of Small Object Detection in Digital Twinning

Specifically, the proposed SOD-DT first integrates the depthwise separable convolution network of MobileNetv2 into YOLOv4 and replaces the original CSPDarknet53, which can provide the rich semantic information for the prediction layer. The fusion of shallow and deep features is then used to increase the accuracy of small target detection. In particular, the human region is extracted as the input for the Openpose-based posture recognition, which may better detect the operators' actions by removing the background interference.

As the basic framework shown in Fig. 2, generally, we mainly focus on two specific modules to realize the intelligent small object detection in digital twinning. First, the depthwise separable convolutions of MobileNetv2 are integrated into YOLOv4, as a newly structured network called YOLOv4-M2, for feature extraction in SOD-DT, which is also used for static small object detections (e.g., equipment, product) in digital twinning. Operations based on depthwise separable convolutions are improved using smaller convolutions to reduce the number of operations and parameters. Furthermore, the generated feature map of YOLOv4 is extended based on the integration of different feature samplings, to facilitate the prediction of small objects. Second, parts of features extracted from YOLOv4-M2 are further fused as input of the Openpose network, which replaces the original VGG-19, so as to save the computing resource and alleviate the gradient disappearance and performance degradation in too deep convolutions. Thus, the improved Openpose network based on feature fusion from shallow and deep layers in the integrated YOLOv4-M2 can reduce the unnecessary background noises, and focus on learning precise human skeleton features to enhance the detection accuracy in long-distance human posture recognition for DT.

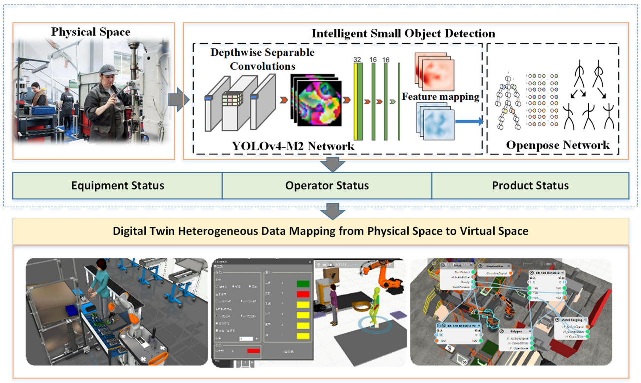

Physical Space
Intelligent Small Object Detection
Depthwise Separable Convolutions
32
16
16
Feature mapping
YOLOv4-M2 Network
Openpose Network
Equipment Status
Operator Status
Product Status
Digital Twin Heterogeneous Data Mapping from Physical Space to Virtual Space

Fig. 2. Framework of intelligent small object detection in digital twinning.

IV. Mechanism of Intelligent Object Detection in Digital Twining

In this section, we discuss the detailed mechanism and implementation of the proposed SOD-DT, including the hybrid neural network architecture, feature fusion based on the integrated YOLOv4-M2, long-distance human posture recognition, and multitype object detection algorithm.

A. Integration of YOLOv4 and MobileNetv2 for Feature Fusion and Extraction

YOLOv4 for object detection is mainly divided into two parts: feature extraction and target prediction. The feature extraction is mainly conducted by the CSPDarknet53 network, in which features represented in each convolution layer are different. The shallow layer contains a large amount of detailed information, such as shapes, textures, and boundaries, which is easily lost after too many convolution and pooling operations. Contrastively, the feature map generated in the deep layer is in smaller size but contains rich semantic information. To improve the accuracy and real-time performance for small object detection, we integrate the MobileNetv2 and YOLOv4 as a new YOLOv4-M2 network in the following way: i) using the depthwise separable convolution network of MobileNetv2 to replace CSPDarknet53

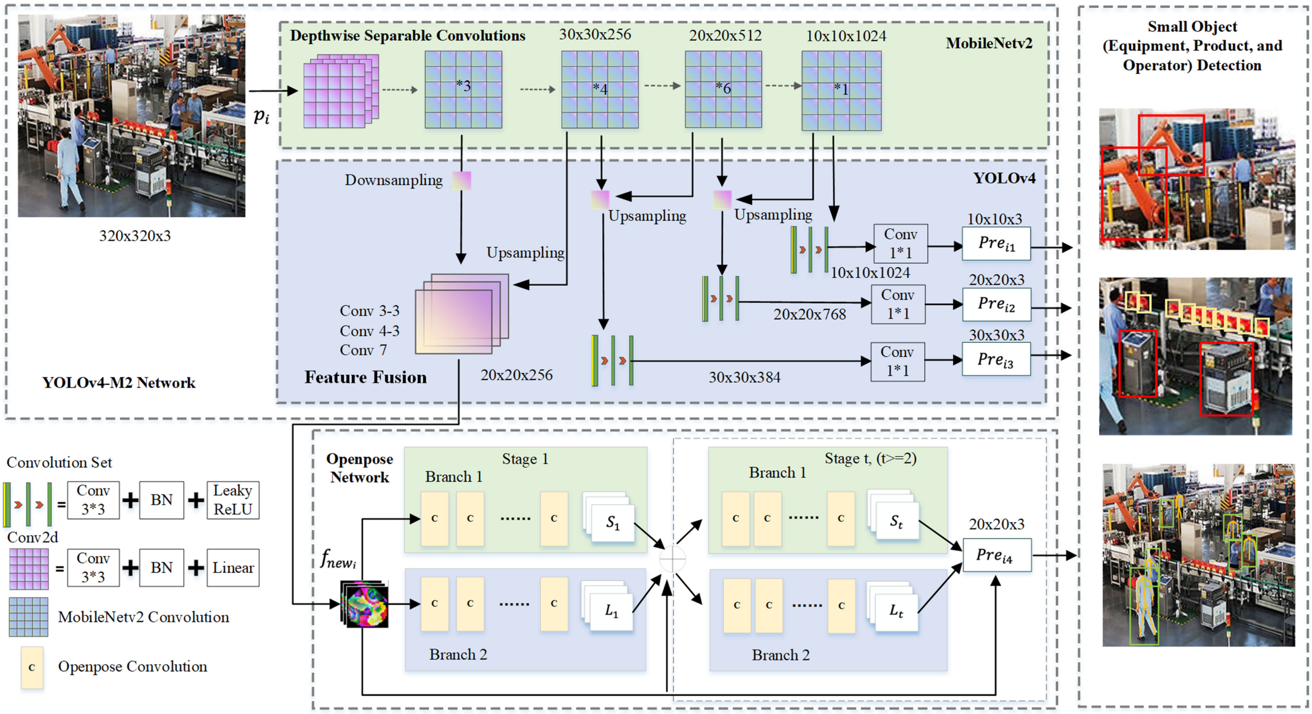


Fig. 3. Integrated neural network architecture for SOD-DT.

TABLE I
INPUT AND OUTPUT OF CONVOLUTIONS

Layer	Input	Convolution	Output
Lyr_{exp}	$H \times W \times D$	1×1	$H \times W \times t \times D$
Lyr_{dep}	$H \times W \times t * D$	$K \times K, stride = S$	$\frac{H}{S} \times \frac{W}{S} \times t \times D$
Lyr_{pro}	$\frac{H}{S} \times \frac{W}{S} \times t * D$	1×1	$\frac{H}{S} \times \frac{W}{S} \times N$

in YOLOv4 for feature extraction, which can efficiently reduce the computational cost; and ii) selecting convolution layers, Conv3-3, Conv4-3, and Conv7, into a new layer with the same channel and pixel, which realizes feature fusion of detail features from the shallow layer and semantic features from the deep layer.

In detail, the depthwise separable convolution network for feature extraction, which is employed to reduce network parameters and convolution operations, is designed with three basic layers. The first layer is the expansion layer Lyr_{exp}. A 1×1 convolution is used to expand the number of channels in the input data. The second layer is the depthwise convolution layer Lyr_{dep}. A 3×3 convolution without pooling layer is used to filter the inputs from the first layer. The third layer is the projection layer Lyr_{pro}. A 1×1 convolution is used to project the high-dimensional data into the low-dimensional one. In addition, the linear activation function is used to alleviate the information loss or even corruption instead of the original ReLU in the first and second layers.

Table I shows the detailed design of input and output in each layer. The input size in the first layer is $H \times W$, and the number of input channels is D, while the number of output channels in the third layer is N. The size of the convolution kernel in the second layer is K, and the stride is S. The expansion factor is t $(0 < t < 1)$. For example, when we set $K = 3$, and $D = N$, the time complexity of this convolution operation can be eight or nine times less than that of the standard convolution.

The detailed structure of the newly designed YOLOv4-M2 is shown in Fig. 3, in which the MobileNetv2 part is used for feature extraction, while the YOLOv4 part is used for object detection. Specifically, the resolution of input data is resized to 320×320, and transformed to a $10 \times 10 \times 1024$ feature map, as the input of YOLOv4 to enhance the further static small target predictions.

Before conducting feature fusion based on the mentioned three convolution layers, Conv3-3, which is located in the shallow layer of YOLOv4-M2, needs to reduce the size, but expand the perceptual field of feature map with key information. The dilated convolution is employed for downsampling during this process, which can be described as follows.

$$S_{\text{in}} = \frac{1}{l}\left[S_{\text{out}} + 2\alpha - r\left(k - 1\right)\right] + 1 \qquad (1)$$

where α is the value of fill pixels, r is the dilation rate, l is the step length, k is the size of convolution kernel. S_{out} is the size of the output feature map, and S_{in} is the size of the input feature map.

Furthermore, feature channels in Conv4-3 and Conv7 need to be compressed, respectively, so as to reduce the number

TABLE II
RESULTS OF FEATURE MAP FUSION

	Sizes of feature map	Number of channels
Conv3-3	Downsampling: 40× 40 →20× 20	256
Conv4-3	Keep as: 20× 20	512 → 256
Conv7	Upsampling: 10× 10 →20× 20	1024 → 256

of parameters and further improve the real-time detection performance. Given an input feature map, the size of which is $H \times W \times D$, the detailed compression operation can be described as follows:

$$C_d = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} E_{dhw} \tag{2}$$

where C_d indicates the output of the compression operation in the *d*th channel, E_{dhw} is the pixel in the *h*th row and *w*th column of the *d*th channel.

After the compression, the activation function used in each channel can be improved and described as follows:

$$\Phi = \sigma\left(M_1\left(\mathcal{L}\left(M_2(z)\right)\right)\right) \tag{3}$$

where z is the output after the compression, $M_1(*)$, $M_2(*)$ are the functions of fully connected layers, $\mathcal{L}$ indicates the linear function, and σ indicates the Sigmoid function.

Based on these, the operation of weight reassignment for each channel can be described as follows:

$$W\left(w_{\text{conv}},\ F_{\text{conv}}\right) = w_{\text{conv}} \cdot F_{\text{conv}} \tag{4}$$

where w_{conv} is the original weight, F_{conv} indicates the feature map after the compression, and $W(*)$ indicates the operation of channel-wise multiplication.

Table II shows the results of feature map fusion of Conv3-3, Conv4-3, and Conv7 in terms of their sizes and numbers of channels. It is noted that the upsampling in Conv7 is conducted based on the standard deconvolution in YOLOv4. Following this way, the extracted feature maps of Conv3-3, Conv4-3, and Conv7 can be seamlessly stitched together as one new feature map, which will be further utilized to enhance the dynamic small target prediction based on long-distance human posture recognition.

B. Long-Distance Human Posture Recognition

The Openpose can be viewed as a parallel convolutional network model, in which one convolutional network works for locating the key points of the human body, while another one is responsible for connecting the candidate key points to form a limb [28]. In addition, the VGG-19 network is used in the original Openpose to extract the features, then feeds them into the parallel convolutional networks, which usually suffer gradient disappearance and performance degradation issues as the number of convolutions increases. Accordingly, as discussed above, a new feature map f_{new} based on the fusion of Conv3-3, Conv4-3, and Conv7 from YOLOv4-M2, is utilized as the extracted features to input into Openpose instead of the original VGG-19, which may efficiently enhance the nonlinear fitting ability of the network, and further improve the accuracy in long-distance recognition.

Algorithm 1: Multitype Object Detection.

```
Input: Frame set $P = \{p_i \,|i = 1, 2, \ldots, n\}$
Output: Prediction result set $Q = \{q_i\}$
Initialize convolution kernel $k = 3$, $Q = \emptyset$;
for each frame $p_i \in P$ do:
    Initialize predict result $q_i = \emptyset$ for $p_i$;
    Resize $p_i$ to the regular pixel 320 × 320;
    Extract the feature $f_i$ using $Lyr_{\text{exp}}$, $Lyr_{\text{dep}}$, and
    $Lyr_{\text{pro}}$ with $k$;
    for $j = 1$ to 3 do:
        Send $f_i$ to YOLOv4 to generate Predict $Pre_{ij}$;
        $q_i = q_i \cup Pre_{ij}$;
    end for
    Transform Conv3-3 to size 20 × 20 by Eq. (1);
    Compress channels in Conv4-3 to size 256 by Eq.
    (2);
    Transform Conv7 to size 20 × 20 by Eq. (1) and
    compress its channels to size 256 by Eq. (2);
    Reactivate each channel by Eq. (3);
    Reassign weights for each channel by Eq. (4);
    Generate the new feature $f_{\text{new}_i}$ as input for
    Openpose;
    Use Openpose to generate Predict $Pre_{i4}$;
    $q_i = q_i \cup Pre_{i4}$;
    $Q = Q \cup q_i$;
end for
return $Q$;
```

As shown in Fig. 3, the whole learning scheme in Openpose can be viewed as a "two-branch and multistage CNN". At Stage 1, Branch 1 produces a set of confidence maps S_1 based on the input f_{new}, which is used to describe the detected human joints, while Branch 2 produces a set of so-called part affinity fields L_1, which is used to assemble the connected joints to predict the human skeleton. In the following each Stage *t*, the input will consist of three parts such as the original f_{new}, and S_{t-1}, L_{t-1} from the previous stage. Predictions from the two branches, S_t and L_t, along with f_{new}, will be associated together for the next stage and finally refine the human posture recognition. During this process, f_{new}, integrating the advantages from the shallow and deep layers, can be further refined to emphasize the skeleton features from the complex large-scale scenes, so as to benefit the long-distance posture recognition in digital twinning.

C. Multitype Object Detection Algorithm

The detailed algorithm for multitype object detection is illustrated in Algorithm 1.

According to Fig. 3, to realize the multitype small object detection in digital twinning, we first resize the resolution of input data to 320 × 320 and feed it into MobileNetv2 for feature extraction. Given an input frame set $P = \{p_i \,|i = 1, 2, \ldots, n\}$, for each p_i, a 10 × 10 feature map can be generated after the depthwise separable convolutions, which is further employed for Predict Pre_{i1}. We then use the upsampling of the 10 × 10 feature map to generate a 20 × 20 feature map, and integrate it with a 20 × 20 feature map from the previous convolution, which is further employed for Predict Pre_{i2}. Likewise, we use the upsampling of the 20 × 20 feature map to generate a 30 × 30 feature map, and integrate it with a 30 × 30 feature map from the previous convolution, which is further employed for Predict Pre_{i3}. Accordingly, the comprehensive results integrated based on Predict Pre_{i1}, Pre_{i2}, and Pre_{i3} using YOLOv4 are utilized for detections of static small objects (e.g., equipment and product) from large-scale scenes in digital twinning. Meanwhile, as a result of feature fusion from the shallow and deep layers in the integrated YOLOv4-M2, a newly generated feature map f_{new_i} based on the integration of convolution layers, Conv3-3, Conv4-3, and Conv7, is used as input for the parallel convolutional network in Openpose. Predict Pre_{i4} is utilized to improve the long-distance human posture recognition, which can enhance the detection of dynamic small objects (e.g., operators) from complex manufacturing environments in digital twinning.

V. Experiment and Analysis

In this section, we evaluate the performance of the proposed model and method for object detection in digital twinning based on three different use cases, comparing with several baseline learning algorithms.

A. Dataset

A real surveillance video dataset was utilized to conduct the evaluation experiment, in which each kind of object samples has nearly 5000 images. We divided the labeled data into 3000 images for training and 2000 images for testing. The size of training images is 320 × 320 × 3. All experiments have been run on a server of Intel Xeon E2288@3.4GHz CPU, 64GB RAM, NVidia GeForce GTX 1080 Ti GPU, Linux, Python 3.7, TensorFlow r2.0.

Three well-known machine learning algorithms, namely, Faster R-CNN, YOLOv3, and SSD, are employed as the baseline methods for performance comparisons. Four widely used metrics, including Precision, Recall, F1, and Accuracy, are employed and calculated for evaluations, according to whether the actual objects have been correctly recognized or not.

B. Learning Performance Comparison

Fig. 4 shows the training efficiency of our proposed SOD-DT comparing with the three baseline methods. X-axis indicates the number of iterations, and Y-axis indicates the values of error rate.

We iterated 400 times to demonstrate the training process. Basically, the error rates of all the four methods decline sharply in the first 50 iterations, and become relatively stable after 150 iterations. Benefitting from the integrated neural network for feature learning and fusion, the proposed SOD-DT obviously outperforms the other three methods, and its error rate fluctuates smoothly after 100 iterations.

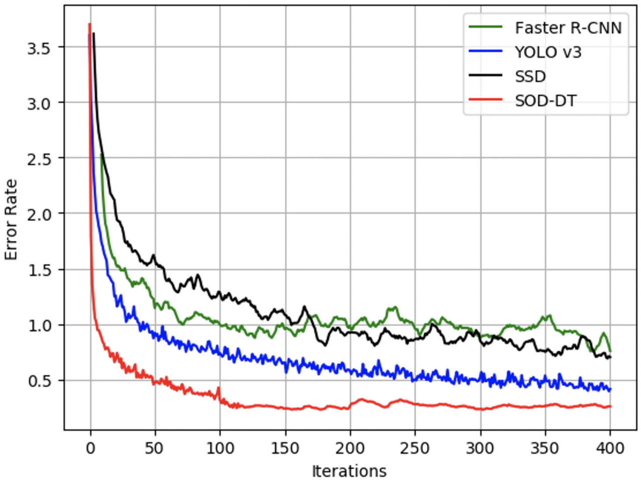


Fig. 4. Training process comparisons of different methods.

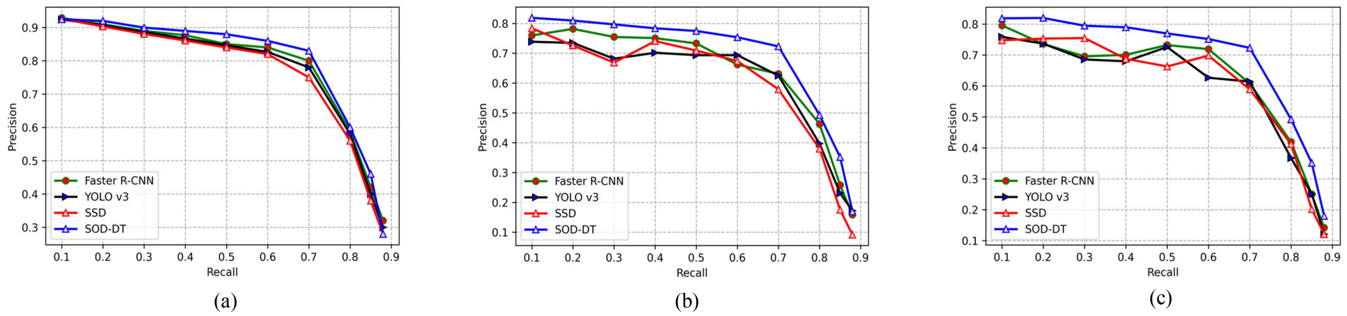
We then demonstrate the performances of four methods in three different cases, considering the distance between the camera and target in real manufacturing environments, i.e., Distance: 8–10 m, Distance: 15–20 m, and Distance: 8–10 m in a blurry environment. The results based on precision–recall curves are shown in Fig. 5.

Fig. 5(a)–(c) shows the performances of four methods in distance between 8 and 10 m, 15 and 20 m, and 8 and10 m in a blurry environment, respectively. It is noted that the camera distance in Fig. 5(c) is as the same as Fig. 5(a), but the light intensity is reduced by 30% to blur the overall environment in the experiment. In general, our SOD-DT performs better than the other three methods in all three cases. Performances of all the four methods basically achieve the same level when the detection distance is relatively close to the object as shown in Fig. 5(a). According to Fig. 5(b) and (c), although the performances of all the four methods degrade in both cases of 15–20 m and 8–10 m blurry environment, due to the longer camera distance, lower resolution image, and even worse manufacturing environment, our SOD-DT degrades slightly and obviously outperforms the other three methods. This result indicates that our method for small object detection is more suitable to tackle the complex scene in DT system, because these scenes may usually result in a certain loss of information of target's features, and drop down the detection performances of conventional learning models.

C. Object Detection Efficiency for DT

We go further to evaluate the practical applicability of the proposed method in some real manufacturing scenes. Table III shows the comparison results in terms of mAP (mean average precision), accuracy, F1, frames per second (FPS), and average detection time (ADT). Both the cases of detection distance in 8–10 m and 15–20 m are applied in the comparison evaluation.

According to the results in 8–10 m, the proposed SOD-DT takes an ADT of 13.9 ms with mAP at 78.2% and accuracy at 91.8%, while Faster R-CNN, YOLOv3, and SSD take 33.6, 24.6, and 36.4 ms, respectively, but result in relatively lower mAP and accuracy. Additionally, when the distance increases to 15–20 m,



(a) (b) (c)

Fig. 5. Comparison performance based on precision–recall curves. (a) Distance: 8–10 m. (b) Distance: 15–20 m. (c) Distance: 8–10 m in a blurry environment.

TABLE III
Object Detection Performance Comparisons Based on Different Metrics

Distance between target and camera	Methods	Mean Average Precision (mAP) (%)	Accuracy (%)	F1 (%)	Real-Time Performance (FPS)	Average Detection Time (ADT) (ms)
8-10m	Faster R-CNN	64.3	74.4	68.3	32	33.6
	SSD	67.7	78.2	71.2	31	36.4
	YOLOv3	71.1	83.4	73.5	44	24.6
	SOD-DT	78.2	91.8	78.9	69	13.9
15-20m	Faster R-CNN	53.7	67.4	62.1	18	58.2
	SSD	52.2	71.9	64.8	21	54.1
	YOLOv3	53.8	72.1	64.7	24	53.8
	SOD-DT	67.1	84.2	72.3	47	29.5

although the ADT of the SOD-DT increases to 29.5 ms, and mAP and accuracy decrease to 67.1% and 84.2%, respectively, the performance is even better than the other three methods in both detection efficiency and accuracy.

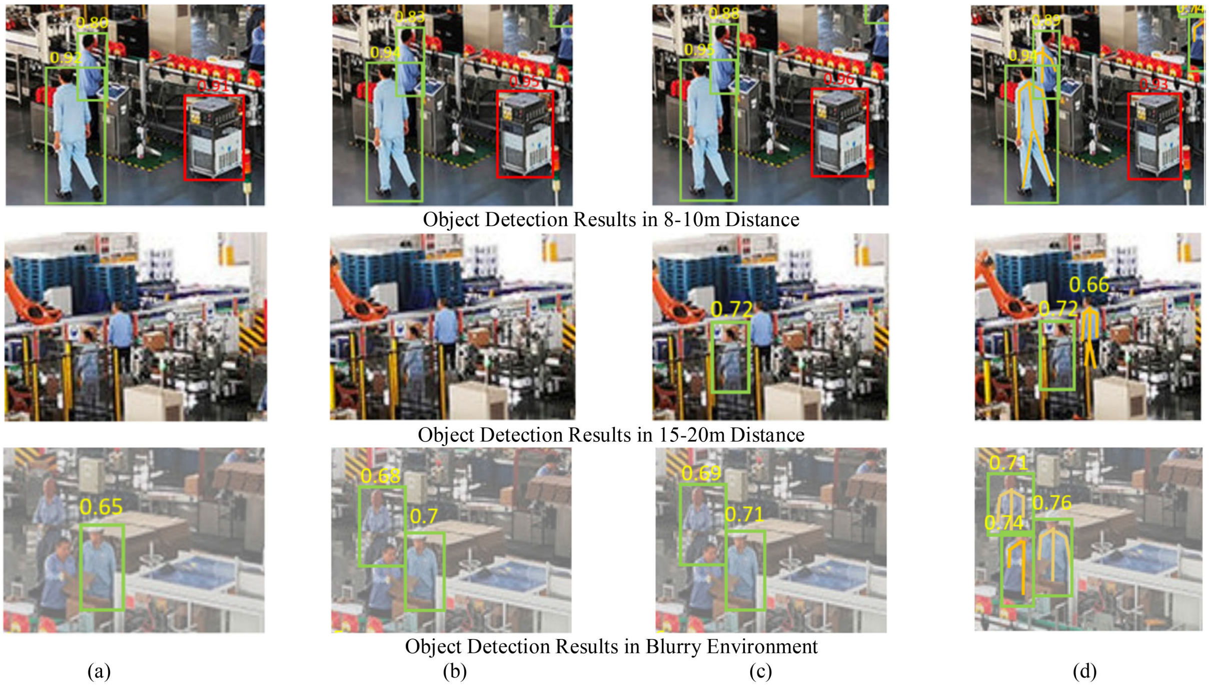
Furthermore, we investigate the performance on real-time status detection of the four methods based on FPS. As shown in Table III, when the detected target is around 8–10 m away from the camera, the proposed SOD-DT improves FPS by 115.6% compared with Faster R-CNN, 56.8% compared with YOLOv3, and 122.6% compared with SSD. When the distance increases to 15–20 m, the results of FPS for all the methods decrease due to the fact that targets become smaller in a relatively more complex environment. It is observed that FPS of Faster R-CNN decreases by 43.8%, YOLOV3 decreases by 45.5%, SSD decreases by 32.3%, and only our SOD-DT decreases by 31.9%. These results indicate that our proposed method can efficiently handle a real-time detection scenario for DT.

Finally, Fig. 6 demonstrates the small object detection evaluation based on the three use cases in a real DT manufacturing environment, in which the objects are mainly composed of equipment devices, products, and human operators. Scores demonstrated in Fig. 6 indicate that fusion of multilevel features based on our hybrid deep neural network can effectively improve the capability of small object detection in complex DT applications.

VI. Conclusion

In this article, to facilitate the modeling and cooperation between a physical manufacturing system and its virtual representation, we investigated the small object detection problem in the complex and large-scale scene of smart manufacturing for DT. A framework of SOD-DT was presented to identify, analyze, and estimate the dynamic changes and real-time status of three important elements: equipment, product, and operator in physical manufacturing space, which could be employed to describe the basic environmental parameters in building a generic DT system of smart manufacturing workshop. A hybrid deep learning model was constructed based on a seamless integration of three neural networks, including MobileNetv2, YOLOv4, and Openpose. Specifically, the depthwise separable convolutions of MobileNetv2 were integrated into YOLOv4 to improve the feature extraction instead of the original CSPDarknet53, which was further utilized to enhance the static small object detections (e.g., equipment and product). The feature map generated from the integrated YOLOv4-M2, instead of the original VGG-19 of Openpose, was then used as input for the parallel convolutional network in Openpose, to enhance the long-distance human posture recognition. Finally, an efficient learning algorithm was developed to realize the multitype small object detection based on the integration and fusion of different feature samplings from shallow and deep layers, which could benefit the modeling, monitoring, and optimizing of the whole manufacturing process in DT system. Experiments and evaluations conducted in three different use cases demonstrated that our proposed SOD-DT was more suitable to cope with the complex situation in large-scale scenes for DT system, compared with three baseline learning algorithms.

In the future, we will further investigate more deep learning schemes to enhance the detection accuracy of multiple objects. More evaluations in different manufacturing scenes will be conducted to improve the model and algorithm with better efficiency.



Object Detection Results in 8-10m Distance

Object Detection Results in 15-20m Distance

Object Detection Results in Blurry Environment

(a) (b) (c) (d)

Fig. 6. Comparisons on object detection for DT among different cases. (a) Faster R-CNN. (b) YOLOv3. (c) SSD. (d) SOD-DT.

References

[1] M. Schluse, M. Priggemeyer, L. Atorf, and J. Rossmann, “Experimentable digital twins—Streamlining simulation-based systems engineering for industry 4.0,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1722–1731, Apr. 2018.

[2] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, “Digital twin-driven product design, manufacturing and service with big data,” *Int. J. Adv. Manuf. Technol.*, vol. 94, pp. 3563–3576, Mar. 2017.

[3] X. Wang, L. T. Yang, Y. Wang, L. Ren, and M. J. Deen, “ADTT: A highly efficient distributed tensor-train decomposition method for IIoT big data,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1573–1582, Mar. 2021.

[4] C. Gehrmann and M. Gunnarsson, “A digital twin based industrial automation and control system security architecture,” *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 669–680, Jan. 2020.

[5] F. Tao and M. Zhang, “Digital twin shop-floor: A new shop-floor paradigm towards smart manufacturing,” *IEEE Access*, vol. 5, pp. 20418–20427, 2017.

[6] Y. Fang, C. Peng, P. Lou, Z. Zhou, J. Hu, and J. Yan, “Digital-twin-based job shop scheduling toward smart manufacturing,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6425–6435, Dec. 2019.

[7] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, “Digital twin in industry: State-of-the-art,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.

[8] A. Rasheed, O. San, and T. Kvamsdal, “Digital twin: Values, challenges and enablers from a modeling perspective,” *IEEE Access*, vol. 8, pp. 21980–22012, 2020.

[9] L. Ren, Z. Meng, X. Wang, L. Zhang, and L. T. Yang, “A data-driven approach of product quality prediction for complex production systems,” *IEEE Trans. Ind. Inform.*, to be published, doi: 10.1109/TII.2020.3001054.

[10] Q. Qi and F. Tao, “Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison,” *IEEE Access*, vol. 6, pp. 3585–3593, 2018.

[11] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and M. J. Deen, “A tensor-based multiattributes visual feature recognition method for industrial intelligence,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2231–2241, Mar. 2021.

[12] M. Schluse, M. Priggemeyer, L. Atorf, and J. Rossmann, “Experimentable digital twins streamlining simulation-based systems engineering for industry 4.0,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1722–1731, Apr. 2018.

[13] C. Zhang, G. Zhou, H. Li, and Y. Cao, “Manufacturing blockchain of things for the configuration of a data- and knowledge-driven digital twin manufacturing cell,” *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11884–11894, Dec. 2020.

[14] J. Zhou, Y. Zhou, B. Wang, and J. Zang, “Human–cyber–physical systems (HCPSs) in the context of new-generation intelligent manufacturing,” *Engineering*, vol. 5, no. 4, pp. 624–636, Aug. 2019.

[15] A. Saad, S. Faddel, T. Youssef, and O. Mohammed, “On the implementation of IoT-based digital twin for networked microgrids resiliency against cyber attacks,” *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5138–5150, Nov. 2020.

[16] Y. Dai, K. Zhang, S. Maharjan, and Y. Zhang, “Deep reinforcement learning for stochastic computation offloading in digital twin networks,” *IEEE Trans. Ind. Inform.*, to be published, doi: 10.1109/TII.2020.3016320.

[17] M. Schluse, M. Priggemeyer, L. Atorf, and J. Rossmann, “Experimentable digital twins—Streamlining simulation-based systems engineering for industry 4.0,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1722–1731, Apr. 2018.

[18] Y. Cai, B. Starly, P. Cohen, and Y.-S. Lee, “Sensor data and information fusion to construct digital-twins virtual machine tools for cyber-physical manufacturing,” *Procedia Manuf.*, vol. 10, pp. 1031–1042, 2017.

[19] J. Leng *et al.*, “ManuChain: Combining permissioned blockchain with a holistic optimization model as bi-level intelligence for smart manufacturing,” *IEEE Trans. Syst. Man Cybern.: Syst.*, vol. 50, no. 1, pp. 182–192, Jan. 2020.

[20] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, “Funnel-structured cascade for multi-view face detection with alignment-awareness,” *Neurocomputing*, vol. 221, pp. 138–145, Jan. 2017.

[21] Y. Jang, H. Gunes, and L. Patras, “Registration-free face-SSD: Single shot analysis of smiles, facial attributes, and affect in the wild,” *Comput. Vis. Image Understanding*, vol. 182, pp. 17–29, May 2019.

[22] 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.

[23] J. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, “Unsupervised object class discovery via saliency-guided multiple class learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.

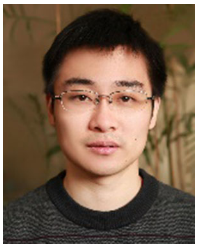
[24] Y. Tang *et al.*, “Visual and semantic knowledge transfer for large scale semi-supervised object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3045–3058, Dec. 2018.

[25] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille, "DeepSkeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5298–5311, Nov. 2017.

[26] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[27] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 712–725, Mar. 2019.

[28] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1302–1310.



Xiaokang Zhou (Member, IEEE) received the Ph.D. degree in human sciences from Waseda University, Tokyo, Japan, in 2014.

He is currently an Associate Professor with the Faculty of Data Science, Shiga University, Hikone, Japan. From 2012 to 2015, he was a Research Associate with the Faculty of Human Sciences, Waseda University. Since 2017, he has also been a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo. He has been engaged in interdisciplinary research works in the fields of computer science and engineering, information systems, and social and human informatics. His research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, and cyber-physical-social system, and cyber intelligence and security.

Dr. Zhou is a Member of the IEEE CS, and ACM USA, IPSJ, and JSAI Japan, and CCF China.

Xuesong Xu (Member, IEEE) received the M.S. and Ph.D. degrees in control science and engineering from Hunan University, Changsha, China, in 2004 and 2009, respectively.

From 2012 to 2015, he was a Postdoctoral Fellow with the National University of Defense and Technology, Changsha. From 2016 to 2017, he was a Visiting Researcher with the Volen National Center for Complex Systems, Brandeis University, USA. He is currently a Professor with the Base of International Science and Technology Innovation and Cooperation on Big Data Technology and Management, Hunan University of Technology and Business, Changsha. He has authored/coauthored more than 30 papers at various conferences and journals, including the *Future Generation Computer Systems*, etc. His research interests include complex system optimization, blockchain, and machine learning.

Dr. Xu is a Member of the IEEE Computational Intelligence, ACM USA and CCF China.

Wei Liang (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2005 and 2016, respectively.

From 2014 to 2015, he was a Researcher with the Department of Human Informatics and Cognitive Sciences, Waseda University, Tokyo, Japan. He is currently with the Base of International Science and Technology Innovation and Cooperation on Big Data Technology and Management, Hunan University of Technology and Business, Changsha. He has authored/coauthored more than 20 papers at various conferences and journals. His research interests include information retrieval, data mining, and artificial intelligence.

Dr. Liang is a Member of the IEEE CS and CCF China.

Zhi Zeng received the B.S. degree in computer science in 2018 from the Hunan University of Technology and Business, Changsha, China, where she is currently working toward the graduate degree in management science and engineering.

Her research interests include artificial intelligence and blockchain.

Shohei Shimizu received the Ph.D. degree in engineering (statistical science) from Osaka University, Suita, Japan, in 2006.

He is currently a Professor with the Faculty of Data Science, Shiga University, Hikone, Japan, and leads the Causal Inference Team, RIKEN Center for Advanced Intelligence Project. His research interests include statistical methodologies for learning data-generating processes, such as structural equation modeling and independent component analysis and their application to causal inference.

Dr. Shimizu received Hayashi Chikio Award (Excellence Award) from the Behaviormetric Society in 2016. Since 2016, he has been a Coordinating Editor of Springer Behaviormetrika.

Laurence T. Yang (Fellow, IEEE) received the B.E. degree in computer science and technology and the B.Sc. degree in applied physics from Tsinghua University, Beijing, China, in 1992, and the Ph.D. degree in computer science from the University of Victoria, BC, Canada, in 2006.

He is currently a Professor and the W.F. James Research Chair with the Department of Computer Science, St. Francis Xavier University, Canada. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data.

Qun Jin (Senior Member, IEEE) received the Ph.D. degree in electrical engineering and computer science from Nihon University, Tokyo, Japan, in 1992.

He is currently a Professor with the Networked Information Systems Laboratory, Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokyo. He has been extensively engaged in research works in the fields of computer science, information systems, and social and human informatics. He seeks to exploit the rich interdependence between theory and practice in his work with interdisciplinary and integrated approaches. His research interests include human-centric ubiquitous computing, behavior and cognitive informatics, big data, data quality assurance and sustainable use, personal analytics and individual modeling, intelligence computing, blockchain, cybersecurity, cyber-enabled applications in healthcare, and computing for well-being.

Dr. Jin is a Senior Member of the Association for Computing Machinery and Information Processing Society of Japan.