Received October 19, 2019, accepted November 10, 2019, date of publication November 26, 2019,
date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955995

3D Object Detection Algorithm for Panoramic Images With Multi-Scale Convolutional Neural Network

DIANWEI WANG[1], (Member, IEEE), YANHUI HE[1], YING LIU[1], (Senior Member, IEEE), DAXIANG LI[1], SHIQIAN WU[2], (Senior Member, IEEE), YONGRUI QIN[3], AND ZHIJIE XU[3], (Member, IEEE)

[1]Center for Image and Information Processing, School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China
[3]School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, U.K.

Corresponding author: Ying Liu (ying.liu.ciip@gmail.com)

This work was supported in part by the Special Project of Foundation Works for Strengthen Police with Science and Technology, Ministry of Public Security, China, under Grant 2019GABJC42, in part by the Dual-Tutor System Project for Innovation and Entrepreneurship of Natural Science Research Program of Shaanxi Province, China, under Grant 2018JM6118, in part by the Innovation and Entrepreneurship Project of Xi'an University of Posts and Telecommunications, China, under Grant 2018SC-08, and in part by the Graduate Innovation Fund of Xi'an University of Posts and Telecommunications, China, under Grant CXJJLA2018005.

ABSTRACT This paper addresses the challenge of 3D object detection from a single panoramic image under severe deformation. The advent of the two-stage approach has impelled significant progress in 3D object detection. However, most available methods only can localize region proposals by a single-scale architecture network, which are sensitive to deformation and distortion. To address this issue, we propose a multi-scale convolutional neural network (MSCNN) to estimate the 3D pose of an object. To be specific, the proposed MSCNN consists of three steps for effectively detecting the distorted object on the panoramic images. The MSCNN contains the CycleGAN network that converts rectilinear images into panoramas, a fused framework that improves both accuracy and speed for object detection, and an adversarial spatial transformer network (ASTN) that extracts the deformation features of the object on panoramic images. Additionally, we recover the 3D pose of the object using a coordinate projection and a 3D bounding box. Extensive experiments demonstrate that the proposed method can achieve a 3D detection accuracy of 38.7% in high-resolution panoramic images, which is higher than the current state-of-the-art algorithm of 5.2%. Moreover, the speed of detection is only about 0.6 seconds per image, which is six times faster than Faster R-CNN (COCO). *The code will be available at https://github.com/Yanhui-He.*

INDEX TERMS Object detection, panoramic images, multi-scale convolutional neural network, 3D bounding box.

I. INTRODUCTION

The panoramic image visualization platform has enjoyed popularity in many applications, such as virtual reality, visual surveillance, autonomous vehicles and virtual interaction [1]. Specifically, the third-generation intelligent video surveillance system and automotive computer vision work have focused on panoramic object detection [2]. Notably, it is important to collect the panoramic image and automatically recognize and detect the objects in them. Panoramic images are typically represented using an equirectangular projection, which creates severe geometric distortions for objects that are further from the central horizontal line [3]. In this projection, the image space coordinates are usually projected onto a focal plane as shown in Fig. 8.

Equirectangular panorama (ERA) can be used to store and transmit VR video. Meanwhile, ERA images create new challenges for computer vision and image processing as i) we lack high-quality annotated 360° datasets, ii) imagery is difficult to treat due to its high-resolution and iii) equirectangular projection creates severe geometric distortions for objects away from the central horizontal line [3]. The distortions of objects vary with distance and viewpoint and reflect randomness to some extent [4]. Therefore, panoramic images create new challenges for object detection, which is a crucial procedure

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

in surveillance videos analysis, industrial inspection, automatic pilot and transportation management [5].

Traditional object detection algorithms are usually based on hand-crafted features, such as histogram of oriented gradient (HOG) [6], local binary pattern (LBP) [7], scale-invariant feature transform (SIFT) [8], frame difference and background subtraction [5], etc. The features of the traditional methods are extracted in the image window and processed with a specific algorithm. However, these sliding-window based approaches can not achieve satisfactory performance of detection because the scales of the objects are always varying significant during moving away from the panoramic camera [4].

The deep learning methods [9]–[14] use CNNs to achieve unprecedented progress in object detection. Compared with traditional features, CNN features are more representative and abstract [4]. Some of these methods have been applied to object detection of panoramic images, especially in 3D object detection in the indoor scene reconstruction, such as PanoContext by Zhang *et al.* [15] and Pano2CAD by Xu *et al.* [16]. The latter retrieves the object by regression using CAD (Computer-Aided Design) models. In contrast, our method does not need any prior knowledge of the object geometry. Moreover, several extended methods [17], [18] have been proposed for 3D object detection, which generated region proposal as an input to the pose network. The pose network is initialized with VGG [19] and fine-tuned for pose estimation using ground truth annotations from Pascal 3D+ [20]. This method is similar to [21] except that the weights and synthetic images are used for training. And some methods are to exploit the availability of 3D shape models and use them for 3D hypothesis sampling and refinement [22]. For example, Mottaghi *et al.* [23] sample the objective viewpoint, position and size and then measure the similarity between rendered 3D CAD models of the object and the detection window using HOG features. Mousavian *et al.* [22] show the 3D pose can be recovered without any template assumptions with carefully-expressed geometrical constraints. Garanderie *et al.* [24] propose a new approach which does not has explicitly expressed geometrical constraints. In this paper, we propose a new method inspired by [24], which has lower computational complexity and higher precision.

The existing work on panoramic images are focusing on indoor scene understanding [15], [16], panoramic to rectilinear video conversion [25], dual camera 360° stereo depth recovery [26], "compression" of wide-angle VR video to conventional narrow-angle video [27], [28], equirectangular super-resolution and 360° object tracking [29]. A wide range of research methods on panoramic vision technology are using active sensing in the form of 360° LIDAR, but it only can perceive object position information. The other method is fusing camera information from multiple different angles [30]. However, the method raises the consumption of computing resources and loses the opportunity of sharing visual information in the early stages of feature extraction due to overlapping fields of view. Furthermore, the image captured from multiple views can also be stitched into a panorama [31], which may lose certain scenes and objects. Compared to LIDAR and multi-view camera fusion, an independent multi-view panoramic camera can provide ultra-high-resolution images in 360° field of view and keep rich scene color and texture information to understand comprehensively on the high-level semantic information [32]. The object in the multi-view panoramic image is deformed and distorted. Object detection tasks have new challenges in this situation. To solve this problem, we propose a new method to detect the objects of 360° panoramic imagery using a multi-scale convolutional neural network (MSCNN).

In terms of geometric transformation and distortion, the features extracted by MSCNN are more robust than the features extracted by convolutional neural network (CNN). Garanderie *et al.* [24] utilized MSCNN to recognize and locate the vehicle in the panoramic image and effectively detected the depth information of the object. However, this approach has a slower detection speed and lower accuracy. In this paper, the proposed method is to improve the detection speed and accuracy by minimizing the network's parameters. Meanwhile, an improved MSCNN is presented for object detection processing in panoramic images. This method combines the region proposal and object detection method to avoid the sliding-window approach altogether and thus it has become adaptive to our application. In addition, we have extended the 2D object detection method based on multi-scale convolutional neural networks. The 3D bounding box of the object is estimated using a method of camera coordinate mapping. Compared with 2D detection, the 3D bounding box can more accurately predict the actual spatial position and attitude of the object, including the coordinates, size, and direction of the object [33]. Alternatively, the 3D object detection methods introduce a third dimension that reveals more detailed object's size and location information [34].

In summary, this work has the following contributions:

- We propose a novel fusion network for 3D object detection with a Multi-scale Convolutional Neural Network (MSCNN), which learns a distortional representation for robust 3D detection localization in panoramic images.
- Based on the predictions from the MSCNN, we propose to utilize an adversarial spatial transformer network (ASTN) to make the whole training process to adapt the panorama.
- We create a new dataset for evaluation on multi-scale panoramic images (4608 × 3456 image resolution) using the CycleGAN network. The experiments show that the proposed approach has a higher accuracy of detection compared to the state of the art on the panoramic datasets using the Darknet-53 framework, and has a much faster learning speed than others in the meanwhile.

II. METHODOLOGY

In this paper, we propose a novel framework for 3D object detection. Given an image, the task of 3D detection is to detect

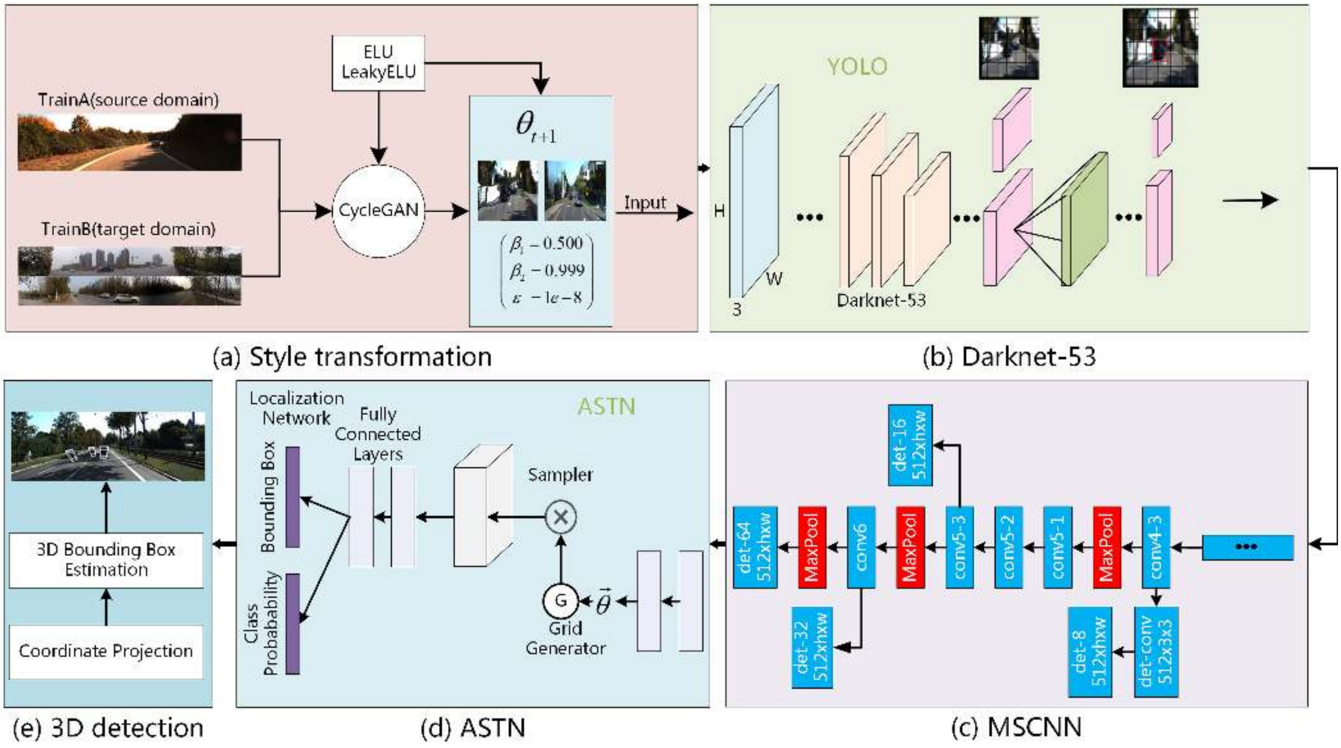

FIGURE 1. Main steps of the 3D object detection based on panoramic images using multi-scale convolutional neural network.

objects and estimate their pose and position. Specifically, 3D detection is represented by a cube in a panoramic image using the Multi-scale Convolutional Neural Network (MSCNN). In a nutshell, we propose the following steps to construct a panoramic 3D object detection algorithm.

Given an image, we utilize a style transformation method to train a model that converts the rectilinear image into a panoramic image. We construct a training dataset of panoramic images and then annotate the ultra-high-resolution image. Subsequently, we reconstruct the first stage of the MSCNN training network on the above dataset by applying the Darknet-53 framework. Specifically, based on an adversarial spatial transformer network (ASTN), we extract deformation features of the object on panoramic images. For each feature, we estimate the parameters of the corresponding 3D bounding box. Finally, we recover the 3D pose of the object using a coordinate projection and a 3D bounding box.

A. TRAINING DATASET GENERATION

Our method can handle multi-scale objects based on the strategy proposed in this section. Due to the high cost of acquiring a significant number of panoramic images and annotating labels to them, we present a training dataset generation method based on CycleGAN [35]. Inspired by the method, we train a transformation model between KITTI [36] rectilinear images and panoramic images. In our approach, the source domain is the KITTI dataset of rectilinear images captured using a front-facing camera rig (82.5° horizontal FoV and 29.7° vertical FoV). Meanwhile, our target domain consists of 5000 panoramic images from a panoramic camera with seven fisheye lenses (360° × 170°). The panoramic image has an angular coverage that is 24.9 times larger than the source KITTI imagery. Subsequently, the transformation model is put forward to transfer all images in the source domain into panoramic style images.

Figure 2 shows some panoramic style images. Most of these generated images have some degrees of distortion and distortion, but that's exactly what we need. Note that, the model used for transformation is only trained on the KITTI and panoramic dataset, where the network adapts to the panoramic image. In the subsequent training model, a large number of parameters related to the real panoramic image style are saved, by which we can detect objects in a real panoramic image.

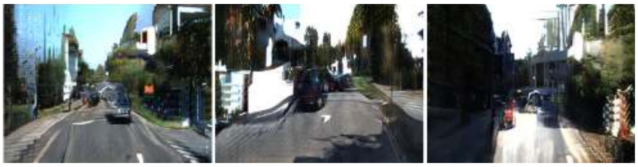

FIGURE 2. Output of the transformation of images from the KITTI dataset.

Implementation: To improve the adaptability of the proposed network to the panoramic image style, we use the following technique to update the parameters,

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon}\hat{m}_t, \tag{1}$$

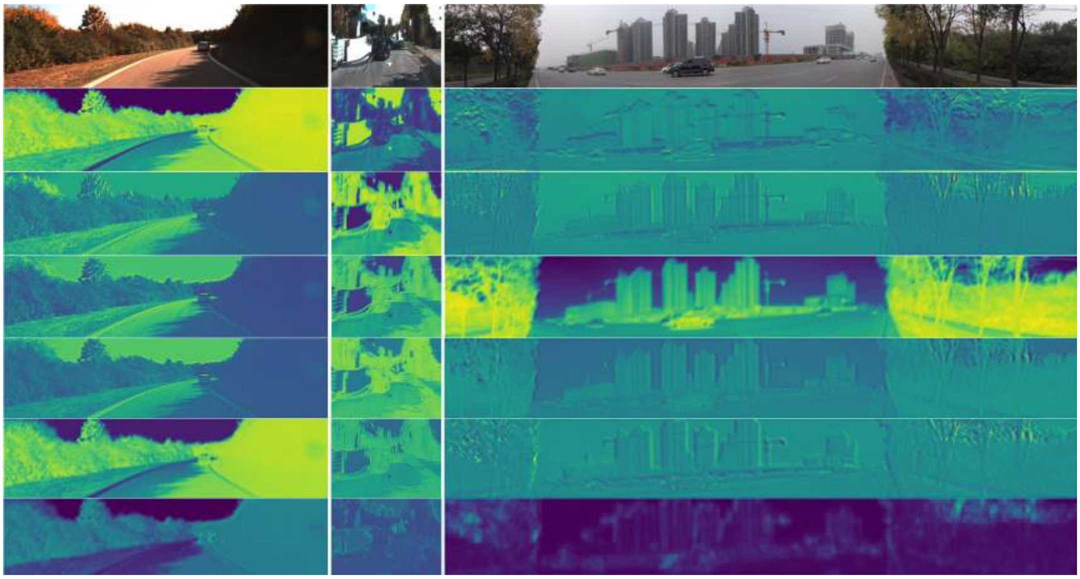

FIGURE 3. Feature maps. From left to right are the outputs from training datasets, the results of KITTI [36], transformed images and real panoramic images.

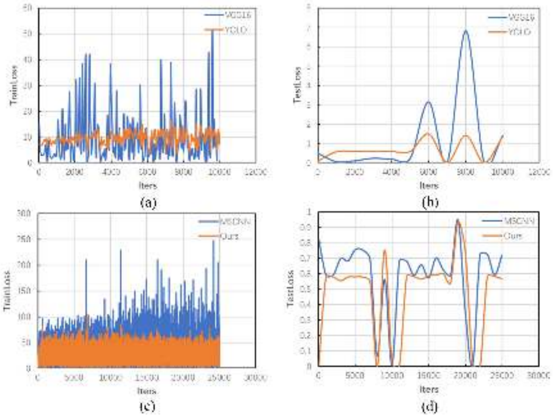

FIGURE 4. Different loss results. (a)-(b): the train and test loss of YOLO. (c)-(d): the train and test loss of MSCNN.

where t denotes the parameter value of the t-th iteration, η is learning rate, ε is a constant used to stabilize the value, $\hat{m}_t$ and $\hat{v}_t$ are the deviation correction for the first and second moment estimates of the gradient respectively in the training process of CycleGAN [35].

B. NETWORK RECONSTRUCTION

The MSCNN is a general multi-class object detector, which is suitable for detecting small-scale objects. We applied this type of network to realize the object classification and detection processing on panoramic images. Figure 1(c) shows the proposed pipeline of MSCNN. The network contains a standard CNN trunk as well as a set of output branches. Furthermore, the network is applicable to detect long-distance objects (the object becomes smaller in this situation) on panoramic images. The training network consists of two stages. In the first stage, the network is trained by the VGG16 [19] model, which adjusts the size of the anchor, and the second stage depends on the result of the previous training stage to provide better initialization.

In order to verify the performance of the style-converted image in network reconstruction, we analyze the output feature map after fine-tuning YOLO [12] on the Darknet-53 framework, as shown in Figure 3. The first column is the KITTI [36] image, the second column is the converted image, and the third column is the real panorama image. We divide the entire YOLO network into five convolutional blocks, each of them consists of 3 to 5 convolutional layers, and all the feature maps of each convolution block are fused in a scale of 1:1 to obtain the fused images that are shown in Figure 3. The Figure 3 indicates the features of the original image, these features are more pronounced in the first and fifth convolution block. The features of the real world panoramic image are prominent on the third convolution block. More specifically, most of the features of the style-converted image change uniformly layer by layer, which reduces the amplitude of the entire network loss and accelerates convergence. Therefore, style-converted images are essential for network training.

Additionally, we investigate the impact of different connections on the feature map of our network. We used a feature map comprised to images of YOLO [12]. The last row of Figure 3 reports an intermediate feature map generated by the MSCNN, while one interesting finding is that the fusion process still retains most of the information of the feature map.

Implementation: First, we construct a Darknet-53 framework on panoramic images without any significant architectural alterations. Then, we use MSCNN instead of the VGG16 [19] to compile a new training network.

C. NETWORK FUSION

This paper focuses on 3D object detection for panoramic images. Given an equirectangular image, the proposed network predicts the pose of objects that are suffering from severe geometric distortions. As for a convolutional neural network (CNN), the fully connected layer can not possess the characteristics of translation invariance in the panoramic. More exactly, the regions of the feature map extracted by CNN will change slightly when the distortions of objects vary with distance and viewpoint. Meanwhile, these errors accumulate continuously when training the network on panoramic images, which leads to poor robust stability of the detection model. For each object, it always has distortions caused by lens distortion and stitching. At run-time, the network is required to have a strong fault-tolerant capability. To avoid the distortion of panoramic images, we utilize the adversarial spatial transformer network (ASTN) as an explicit processing module to extract the features from CNN, which can independently learn the translation and rotation invariance.

Implementation: We add the ASTN network as an input after the ROI-pooling layer of MSCNN. Panoramic images are deformed and divided into four blocks by the ASTN. For each block, we estimate four different rotation angles.

D. 3D BOUNDING BOX GENERATION

We train a multi-scale convolutional neural network for detecting objects in panoramic images, which helps us to handle small object localization errors made during detection. The model includes three types of information: position, category, and spatial features. Here, we use coordinate projection to emphasize the geometric relationship between the rectilinear image and the panorama. Moreover, we utilize the bounding box to estimate the spatial feature of the object. We detail these steps in the sub-sections below.

1) COORDINATE PROJECTION

In order to project the spatial coordinates of the panoramic image to the plane, we provided the Cartesian coordinates of a 3D scene point in camera space. To simplify the calculation, the longitude and latitude are defined as:

$$(\lambda, \phi)^{-1} = \Gamma\left[(x, y, z)^{-1}\right] = \Gamma\left[(\alpha, \beta, 1)^{-1}\right], \qquad (2)$$

where $\alpha = x/z$ and $\beta = y/z$.

We define an image transformation matrix T_p which transforms the longitude and latitude to image space coordinates (u_p,v_p):

$$\begin{bmatrix} u_p \\ v_p \\ 1 \end{bmatrix} = T_p \cdot \begin{bmatrix} \lambda \\ \phi \\ 1 \end{bmatrix} = \begin{bmatrix} \gamma & 0 & c_\lambda \\ 0 & \gamma & c_\phi \\ 0 & 0 & 1 \end{bmatrix} \cdot \Gamma\left(\begin{bmatrix} \alpha \\ \beta \\ 1 \end{bmatrix}\right), \qquad (3)$$

where γ and (c_λ,c_ϕ) are the angular resolution parameter and the principal point of the camera respectively. The equirectangular imagery generated by a panoramic camera with seven fisheye cameras can be readily used without any prior calibration.

2) 3D BOUNDING BOX ESTIMATION

In order to estimate the 3D bounding box accurately and robustly, we present an improved method based on MSCNN. We remove the final fully connected MSCNN layer and then regress the direction of objects by the last fully connected ASTN layer. In contrast to directly regressing object locations from the 3D CAD models [15], [16], we can recover the object's 3D bounding box using the constraints of the 2D detection window. Specifically, we directly regress the 3D dimensions (width, length, and height). As a result, we are able to recover the actual 3D position from the panorama by:

$$(x, y, z)^{-1} = r \cdot u\left[\Gamma^{-1}\left[T_p^{-1} \cdot \left(u_p, v_p, 1\right)^{-1}\right]\right]. \qquad (4)$$

Implementation: Here, the angular resolution γ is defined as $\gamma = 2\pi/w$, where w is the width of the image. To simplify the computation, r is set as $r = \gamma h$, where h is the height of the region proposal generated by the Region Proposal Network (RPN). The operation $u[\cdot]$ is defined as $u[\phi] = \phi/\|\phi\|$.

The proposed framework of 3D object detection model for panoramic images is shown in Figure 1. The Fig. 1(a) is the technique to generate the dataset for training by CycleGAN [35]. The Fig. 1(b) is the method to generate the pre-trained model by Darknet-53. The Fig. 1(c)-(d) are the approach to generate the object detection model by MSCNN. The Fig. 1(e) is the method to estimate the parameters of the 3D bounding box. We combine different structural network frameworks by adopting a multi-stage training strategy. CycleGAN is used to enhance the training set to achieve conversion between different data source domains. Subsequently, we fine-tune the YOLO [12] network on the DarkNet-53 to train the network. In the next phase, the ASTN module is added to the MSCNN to improve the robustness of the model to detect objects on the panoramic dataset. Finally, the 3D bounding box of the object is estimated by the coordinate map.

III. EVALUATION

Some technical details of the training network:

1. In the training process of CycleGAN [35], we set the initial learning rate to 0.0002, the parameter of the first moment and second moment are respectively 0.5 and 0.999, ε uses the default value of 1e-8.

2. In the first stage of MSCNN, we make use of the framework of Darknet-53 to train a new model. To prevent overfitting, we set the momentum and the weight decay to 0.9 and 0.005. The initial learning rate is set to 0.001 and a multi-distribution strategy was employed.

3. In the second stage of training, we joined ASTN network. We set the initial learning rate to 0.0005, the momentum and the weight decay are respectively 0.9 and 0.0005.

Training platforms are as follows: GPU Titan XP ×4, CUDA9.0, CUDNN7.0, Ubuntu16.04. The following sections are the results of training and evaluation.

A. DATASETS

1) CARLA

We evaluate our approach both qualitatively and quantitatively on panoramic images from using synthetic data

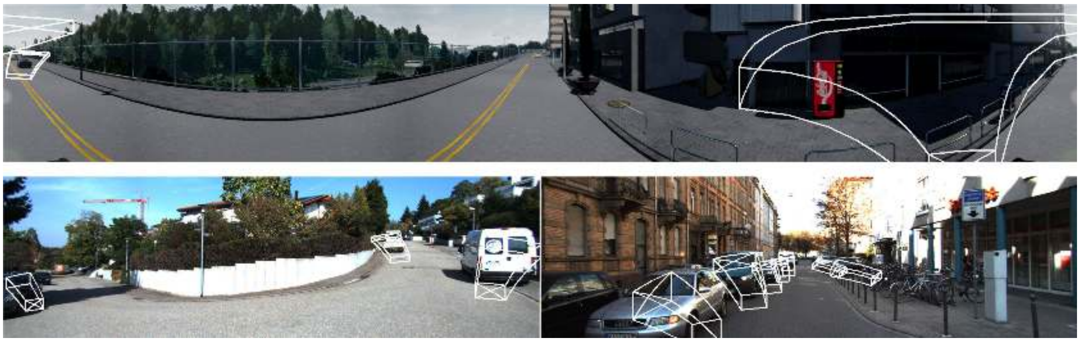

FIGURE 5. 3D object detection results. (Use only the data generated by CycleGAN.)

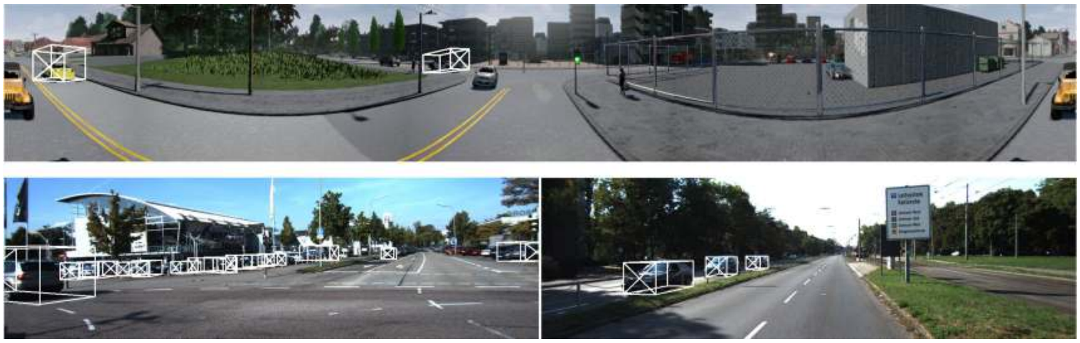

FIGURE 6. 3D object detection results. (Only the fusion framework of Darknet-53 and MSCNN is used.)

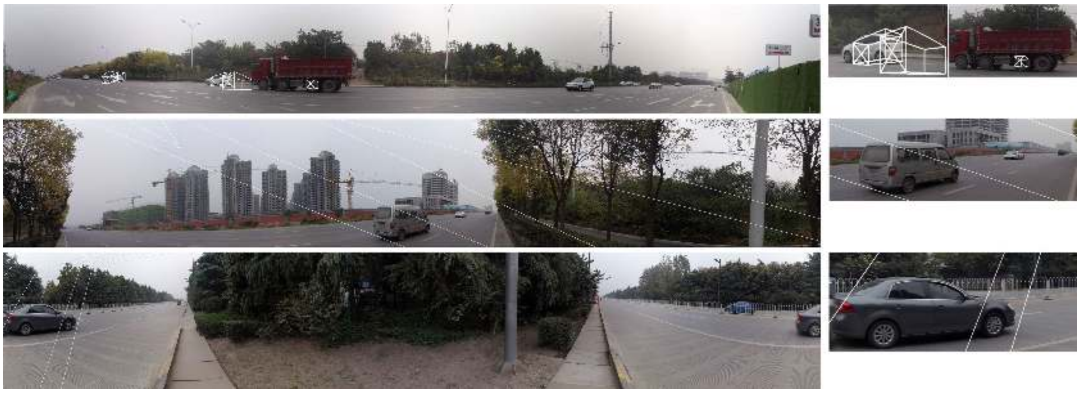

FIGURE 7. 3D object detection results. (Use only ASTN networks.)

generated using the CARLA [37] automotive environment simulator. A total of 200 panoramic simulation images (2048 × 300 image resolution). It is represented by C in Table 1.

2) KITTI

KITTI [36] is the current computer vision algorithm evaluation dataset under the largest autopilot scene in the world. We utilize 7482 images (1242 × 375 image resolution) as the data source for the style conversion image, and another 7519 images for the test. It is represented by K1 in Table 1.

3) PANORAMA

The panorama is a real-world panoramic image captured by a 7-mesh panoramic camera. We evaluate our method with 5,000 of the panoramic images (4608 × 3456 image resolution). It is represented by P in Table 1.

B. TRAINING RESULTS

The converting results from the KITTI [36] image style to the panoramic using the CycleGAN [35] model are shown in Fig. 2. Compared with the original images, the scale transformation of the object is significantly increased. However, the converted image retains the texture and color information on objects very well.

Some loss results are shown in Figure 4. In the first stage of the MSCNN, we find that the training loss based on the Darknet-53 is much smaller compared with VGG16 [19]. Meanwhile, the amplitude of the loss is smaller and the convergence speed is faster than the original network. Especially we found that even though the final test loss results are almost the same during the test, the intermediate process of loss is much stable than the VGG16-based network. In the second stage of the MSCNN, the net loss of joining ASTN is significantly smaller. When the number of iterations

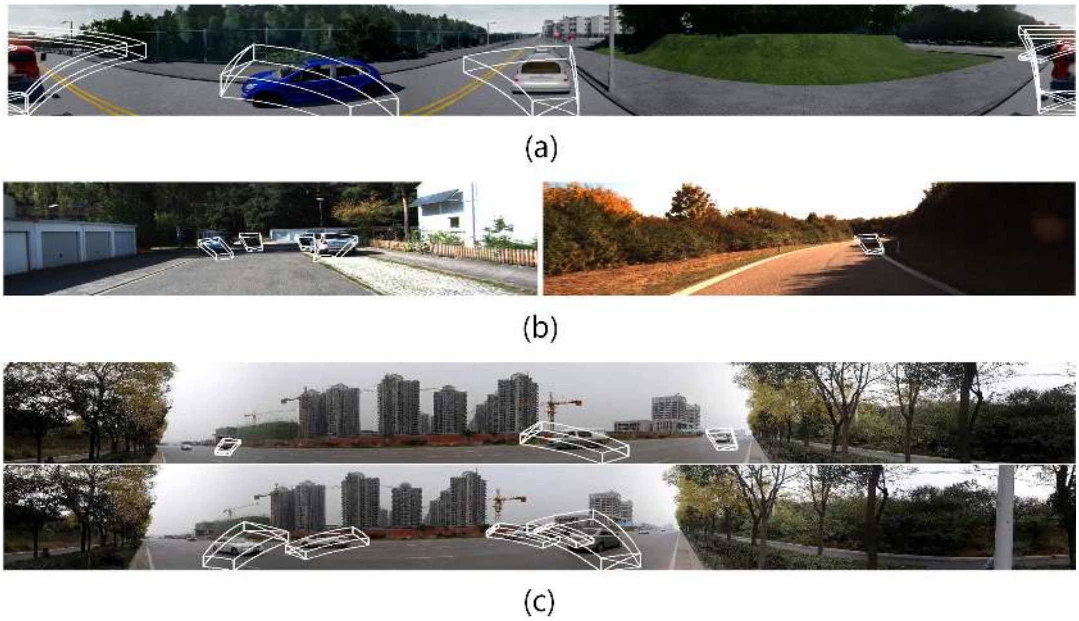

FIGURE 8. 3D object detection results. (a) Synthetic images. (b) Transformed images. (c) Real-world images.

TABLE 1. The accuracies of our method and the time consumption (ms) on the different datasets.

Network	Resolution	Train	mAp	Time
MSCNN&VGG16 [24]	-	C	35.5	-
	-	K1+C	33.5	640
MSCNN&Darknet	715 × 256	K1+C	34.7	580
	715 × 256	K1+P	33.8	585
	715 × 256	K1+P+C	**36.4**	**565**
	1242 × 375	K1+C	35.3	575
	1242 × 375	K1+P	35.0	575
	1242 × 375	K1+P+C	35.3	575
MSCNN&ASTN	715 × 256	K1+C	36.5	580
	715 × 256	K1+P	35.2	580
	715 × 256	K1+P+C	**38.7**	560
	1242 × 375	K1+C	36.0	575
	1242 × 375	K1+P	35.8	560
	1242 × 375	K1+P+C	35.9	**555**
MSCNN&Darknet&ASTN	715 × 256	K1+C	35.4	**565**
	715 × 256	K1+P	36.0	565
	715 × 256	K1+P+C	**37.5**	575
	1242 × 375	K1+C	32.6	595
	1242 × 375	K1+P	34.4	580
	1242 × 375	K1+P+C	36.8	570

is less than 5000, the loss value of the proposed method has stabilized. The final loss value is much smaller than the unimproved MSCNN. The test loss of the entire network is less than 0.6. These results indicate that the performance of the proposed network is better than the original network.

C. QUALITATIVE EVALUATION

Our approach is evaluated on panoramic images by the automotive environment simulator. Meanwhile, we also give the detection results of a certain module in the proposed method (Fig 5.-Fig 7.). Figure 5 shows the detection results on the simulated data and the KITTI dataset. This model is trained only with data generated by CycleGAN. It is indicated that the 3D bounding box has caused a serious deformation in the target area of the object, and marked the object incorrectly. Figure 6 is the detection results on simulated data and the KITTI dataset by the fusion network in our proposed method. The improved model demonstrates that the proposed method has robust detection performance on the KITTI, but the accuracy on the panoramic dataset is not ideal. Figure 7 shows the result of our detection on a real panoramic image by the ASTN network. The model has large errors and false detection. In specific the entire map is completely disrupted when the object scale changes seriously, where the model does not provide enough information to infer the poses. It is indicated that the single ASTN network can not adapt to panoramic images without the training of multi-scale networks.

Some qualitative results are shown in Figure 8. It illustrates that the 3D object detection results of the images in representative scenes. The proposed method can detect the target that appears at the edge of the image in Fig 8(a), which demonstrates that our method enables the model to learn the relationship between parts of the object. For a transformed image, the model can wrap small objects in small cuboids for 3D detection. Even the image background is complex in a real panoramic image, our method robustly detects the object. Some examples are shown in Figure 8(c). Meanwhile, the 3D boxes are approximately enwrapping the object with a white cuboid.

Note that the white 3D bounding box is always not regular cubes in the results. A real panoramic image consist of multiple fisheye images, and the scale of the target varies with distance continuously within a 360-degree horizontal field of view. The improved performance demonstrates that the proposed method enables our network to detect the object in a panorama so that the 3D pose can be robustly recovered by the coordinate mapping.

D. QUANTITATIVE EVALUATION

Table 1 indicates the mean average precision (MAP) and the average time consumption of each image. To validate the validity of the algorithm, we retrain our models with different datasets and the object detection performance is shown in Table 1.

Compared with the method proposed in [24], we only fine-tune the YOLO network on the Darknet-53 deep learning framework, and the accuracy of obtaining the 3D bounding box of the object by coordinate mapping technique is increased by 2.9% (image resolution is 715 × 256) and 1.8% (image resolution is 1242 × 375), and the detection time of each image is reduced by 75 milliseconds. If the ASTN network is cascaded only based on MSCNN, the detection accuracy is increased by 5.2% (image resolution is 715 × 256) and 2.4% (image resolution is 1242 × 375), and the detection time is reduced by 85 milliseconds. Finally, we integrated the YOLO and ASTN networks in the MSCNN network, with detection accuracy increased by 4.0% (image resolution is 715 × 256) and 3.3% (image resolution is 1242 × 375), and detection time reduced by 75 milliseconds.

We also find that the accuracy and speed of the detection do not change when using only the YOLO framework (image resolution is 1242 × 375, the accuracy is about 35% and the speed is about 575 milliseconds per image). To validate the benefit of considering the YOLO in solving the 3D detection problem, we report the extracted feature maps of the Darknet-53 in Figure 3. It is indicated that in high-resolution panoramic images, the learning framework based on multi-scale convolutional neural networks will extract similar features with little effect on the final results. Also, Table 1 shows that the detection accuracy can be improved significantly by training the network with CARLA [37] dataset.

Figure 9. shows the experiment MAP results with comparisons of different resolutions and datasets. We make the following conclusions for each step of the method we propose. The accuracy of network reconstruction using the Darknet-53 framework alone increased by 1.2% (image resolution is 715 × 256) and 1.8% (image resolution is 1242 × 375), respectively. The accuracy of the ASTN network alone was increased by 3.0% (image resolution is 715 × 256) and 2.5% (image resolution is 1242 × 375), respectively. Both methods reduce the detection time by 65 milliseconds (ms). Although the MAP obtained after integrating all the methods is not the highest, the final fusion algorithm performs best in high-resolution panoramic image testing (36.8%).

Our best results are derived from the combined training dataset consisting of KITTI, CARLA, and panoramic images obtained through a panoramic camera with seven fisheye cameras. Table 1 shows the comparison of our methods with the method [24] in terms of MAP metric. Note that the best detection result of the proposed algorithm has an improvement of 5.2% compared to the original network. Mainly attribute to the ASTN enhances the robustness of the network. Meanwhile, the average detection time is reduced by 80 ms. Moreover, the time consumption is directly reduced by 75 ms when we reconstruct the framework with Darknet-53. Even though the resolution of the best images is very high, the detection speed of the proposed network is 6 times faster than the Faster R-CNN [10]. Additionally, the experiment results indicate that the fusion network (pre-trained by the MSCNN and ASTN) has better performance than that of the MSCNN network.

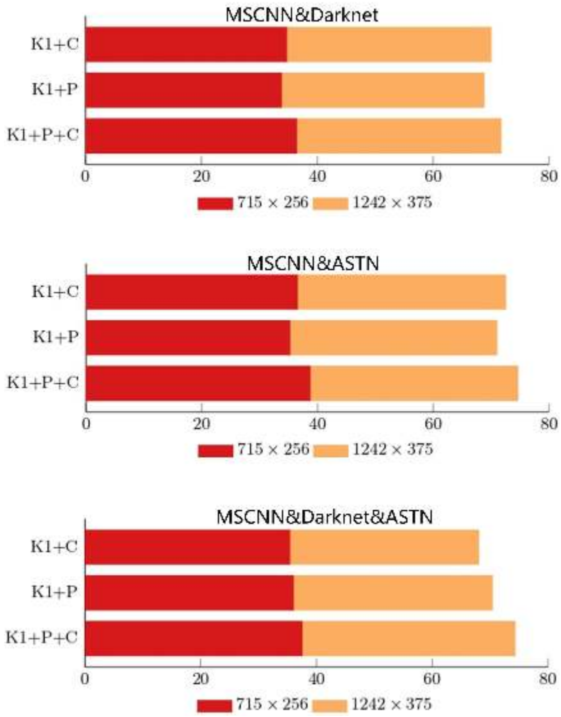

FIGURE 9. **Comparison of experimental results of different methods MAP.**

IV. CONCLUSION

In this paper, we proposed a novel framework for 3D object detection in panoramic images, which consists of the panorama style transformation for training dataset generation and the fusion network for final 3D object detection. We showed that pre-training the network with the transformed image followed by the fusion network for 3D detection gained a superior performance than direct regression of 3D detection by the single MSCNN, especially for multi-scale objects. We also showed that considering the 3D bounding box of predicted object location in the high-resolution image further improved 3D detection. We reported the state-of-the-art performances on different datasets and demonstrated the robustness of the proposed approach on the real panoramic dataset.

REFERENCES

[1] H. Kim, J. Jung, and J. Paik, "Fisheye lens camera based surveillance system for wide field of view monitoring," *Optik*, vol. 127, no. 14, pp. 5636–5646, 2016.

[2] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5420–5428.

[3] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2190–2195.

[4] F. Deng, X. Zhu, and J. Ren, "Object detection on panoramic images based on deep learning," in *Proc. 3rd Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2017, pp. 375–380.

[5] H. Mannila and P. Orponen, *Algorithms and Applications*. Springer, 2010.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[7] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2000, pp. 404–420.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

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

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

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

[14] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[15] Y. Zhang, S. Song, P. Tan, and J. Xiao, "PanoContext: A whole-room 3D context model for panoramic scene understanding," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 668–686.

[16] J. Xu, B. Stenger, T. Kerola, and T. Tung, "Pano2CAD: Room layout from a single panorama image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 354–362.

[17] S. Tulsiani and J. Malik, "Viewpoints and keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1510–1519.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[20] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 649–666.

[21] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2686–2694.

[22] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7074–7082.

[23] R. Mottaghi, Y. Xiang, and S. Savarese, "A coarse-to-fine model for 3D pose estimation and sub-category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 418–426.

[24] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360° panoramic imagery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 789–807.

[25] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1396–1405.

[26] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski, "Low-cost 360 stereo photography and video capture," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 148:1–148:12, Jul. 2017.

[27] W. Bares, V. Gandhi, Q. Galvane, and R. Ronfard, "Pano2Vid: Automatic cinematography for watching 360° videos," in *Proc. Eurograph. Workshop Intell. Cinematogr. Editing*, 2017, p. 1.

[28] Y.-C. Su and K. Grauman, "Making 360 video watchable in 2D: Learning videography for click free viewing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1368–1376.

[29] U. Kart, J.-K. Kämäräinen, L. Fan, and M. Gabbouj, "Evaluation of visual object trackers on equirectangular panorama," in *Proc. VISIGRAPP*, 2018, pp. 25–32.

[30] O. K. Hamilton and T. P. Breckon, "Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3439–3443.

[31] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," in *Proc. CVPR*, Jun. 2005, pp. 510–517.

[32] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," 2017, *arXiv:1704.05519*. [Online]. Available: https://arxiv.org/abs/1704.05519

[33] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, "3D object detection: Learning 3D bounding boxes from scaled down 2D bounding boxes in RGB-D images," *Inf. Sci.*, vol. 476, pp. 147–158, Feb. 2019.

[34] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[37] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," 2017, *arXiv:1711.03938*. [Online]. Available: https://arxiv.org/abs/1711.03938

DIANWEI WANG (M'18) received the B.E. degree in mechanical and electronic engineering from the Gansu University of Science and Technology, in 2002, and the master's degree in mechanical design from Xidian University, in 2005, and the Ph.D. degree in navigation, guidance, and control from Northwestern Polytechnical University, in 2010. From May 2015 to 2016, he was a Visiting Scholar with the University of Huddersfield, U.K. He is currently an Associate Professor with the School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China. He has published over 40 peer-reviewed journal and conference papers as well as two edited books in the relevant fields. His main research interests include image enhancement processing, object detection, target identification and tracking, nonstationary signal processing and analysis, and application of artificial intelligence in the machine vision field. He is a member of CCF and ACM.

YANHUI HE received the B.E. degree from the College of Communication Engineering, Xi'an Technological University, Xi'an, China. He is currently pursuing the master's degree with the School of Communication and Information Technology, Xi'an University of Posts and Telecommunications, Xi'an, where he is also a dual Tutor Graduate Student. His research interest includes image processing via both convolutional and deep learning methods.

YING LIU (M'01–SM'07) received the B.Sc. degree from the School of Information Engineering, Xidian University, China, the M.Eng. degree from the School of Electrical Engineering, National University of Singapore, and the Ph.D. degree with the School of Computing and Information Technology, Monash University, Australia. She is currently a Full Professor with the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications (XUPT), China. She also serves as the Director of the Center for Image and Information Processing, XUPT.

DAXIANG LI received the M.S. degree from the Department of Electronics Science, Northwest University, China, and the Ph.D. degree with the School of Information Science and Technology, Northwest University, in 2010. He has published more than 30 refereed journal and conference papers as well as three edited books in the relevant fields. His main research interests include image retrieval, image classification, and image annotation.

SHIQIAN WU (M'02–SM'05) received the B.Eng. and M.Eng. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1985 and 1988, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2001. He was an Assistant Professor, a Lecturer, and an Associate Professor with HUST, from 1988 to 1997. From 2000 to 2014, he was a Research Fellow and then a Research Scientist with the Agency for Science, Technology and Research, Singapore. He is currently a Professor with the School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan. He has coauthored the book *Dynamic Fuzzy Neural Networks* (Singapore: McGraw-Hill, 2003). He has authored or coauthored over 180 scientific publications (book chapters and journal/conference papers). He was listed as one of the most cited Chinese researchers by Elsevier, in 2017. His current research interests include image processing, pattern recognition, machine vision, fuzzy systems, and neural networks.

YONGRUI QIN received the Ph.D. degree in computer science from The University of Adelaide, Adelaide, Australia, in 2015. He is currently a Senior Lecturer with the School of Computing and Engineering, University of Huddersfield, U.K. He has published more than 80 refereed technical articles, including publications in prestigious journals, such as *ACM Computing Surveys*, the IEEE Transactions on Parallel and Distributed Systems, *ACM Transactions on Internet Technology*, *World Wide Web Journal*, *Journal of Network and Computer Applications*, and the IEEE Internet Computing, and reputable international conferences, such as SIGIR, EDBT, CIKM, CAiSE, WISE, ICWS, SSDBM, and DASFAA. His main research interests include the Internet of Things, graph data management, data stream processing, data mining, information retrieval, semantic web, computer networks, and mobile computing.

ZHIJIE XU (M'07) is currently a Full Professor of visual computing with the University of Huddersfield, U.K. In the last 25 years, his research has been mainly focused on the areas of computational geometry, real-time graphics, and vision systems. He has published over 100 peer-reviewed journal and conference papers as well as five edited books in the relevant fields. He is a Professional Member of IET and BCS and a Fellow of the Higher Education. In addition to his academic roles, such as a Project Leader, a Journal Editor, and the Conference Chair. He has also been actively involved in industrial consultation and government advisory for SGI-focused planning. He has also delivered speeches on various high-tech events to encourage public understanding, participation and debate on digital technology in society.