IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Perspective Transformation Data Augmentation for Object Detection

KE WANG[1], BIN FANG[1], (Senior Member, IEEE) JIYE QIAN[2], SU YANG[1], AND XIN ZHOU.[1] JIE ZHOU[3]

[1]College of Computer Science, Chongqing University, Chongqing, China (e-mail: wangkecqu@cqu.edu.cn)
[2]State Grid Chongqing Electric Power Co. Electric Power Research Institute, Chongqing, China(e-mail: qianjiye@cqu.edu.cn)
[3]State Grid Chongqing Yongchuan Power Supply Company, Chongqing, China(e-mail: zhj_yc@foxmail.com)

Corresponding author: Bin Fang (e-mail: fb@cqu.edu.cn).

This research was supported by the National Natural Science Foundation of China (NO.61876026, NO.61906022, NO.61472053, NO.91420102, NO.61703062) and the Special Foundation for Chongqing Postdoctoral Research (Xm2016060)

ABSTRACT One major reason for the success of convolutional neural networks (CNNs) is the availability of large-scale labeled data. Effective training of CNNs relies on large annotated data. Unfortunately, large amounts of data with corresponding annotations are too expensive to obtain in some real-world applications. One reasonable alternative is to use data augmentation techniques to automatically generate annotated samples. In this paper, a novel data augmentation framework based on perspective transformation is proposed. This method automatically generates new annotated data without extra manual labeling, thus effectively extends the inadequate dataset. Perspective transformation can produce new images captured from any cameras viewpoints. Therefore, our method can mimic images taken at the angle that the camera cannot reach. Extensive experimental results on several datasets have demonstrated that our perspective transformation data augmentation strategy is an effective tool when using deep CNNs on small or imbalance datasets.

INDEX TERMS data augmentation, perspective transformation, object detection, transmission-line

I. INTRODUCTION

OVER recent years, great changes have taken place in the field of computer vision, which is promoted by Convolutional Neural Networks (CNNs) [1]. CNN-based models have achieved state-of-art performance in various tasks, like image classification [2] [3] [4], multi-object detection [5] [6] [7], semantic segmentation [8] [9] [10], and active recognition tasks [11] [12]. The success of CNNs benefits from access to a large amounts of labeled data – a model's performance typically improves as increasing the quality, diversity and the amount of training data. On the one hand, the performance of the model increases logarithmically based on the volume of training data size [13]; on the other hand, even the use of over-fitting prevention techniques [14] [15], using small or imbalanced dataset for training still results in serious over-fitting [16].

However, it is prohibitively difficult to collect enough quality data to train a model for good performance. One way of acquiring large annotated data is to automatically enlarge dataset by existing data information, which is known as data augmentation (DA). Over the years, various data augmentation methods have been proposed as a solution for training CNNs when only limited data is available. In the field of image processing, common augmentation methods include flipping, color shifting, adding noise and so on [2] [6] [17]. Recent work has shown that strategies learned from data can further improve the performance of image classification models [18] [19]. In addition, Generative Adversarial Network can also be used to augment data in image classification task [20] [21].

In multi-object detection tasks, simple augmentation techniques such as flipping, cropping and rotation are usually used. GAN and other learned method cannot be directly applied to detection tasks, because they only produce new images without corresponding annotation files. This is particularly the case for transmission-line object detection. On the one hand, captured by unmanned aerial vehicles (UAVs), transmission-line images are not easily collected due to the UAV limitation and environmental complexity. The scarcity of data results in a lack of samples for rarer but important classes such as composite insulator and glass insulator, which can make it difficult for models to understand these con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

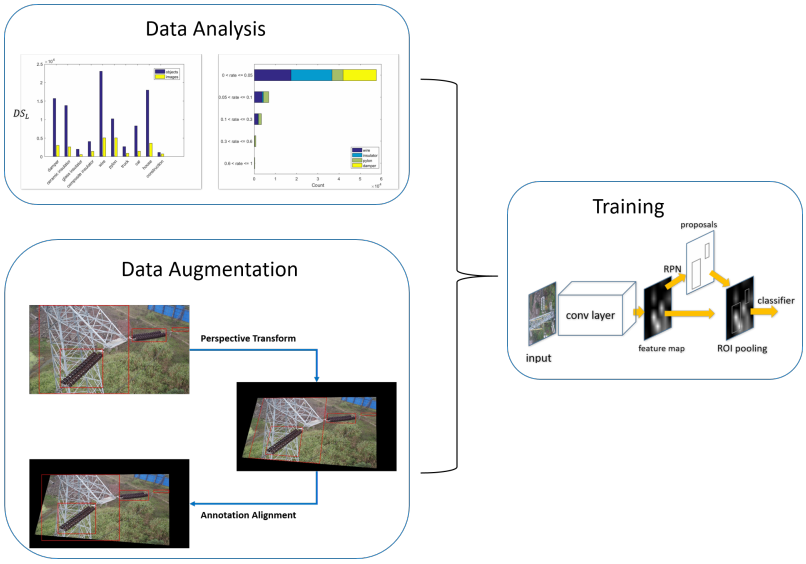


FIGURE 1. An overview of our data augmentation for object detection.

cepts without over-fitting. On the other hand, the collected transmission-line image contains a large number of instance objects (18 instance objects per image), and these instance objects are unevenly distributed. This leads to the manual annotation of transmission-line images is very expensive.

In this work, we use perspective transformation to mimic images taken by camera from different angles. When human eyes see nearby objects, they look bigger than distant objects, which is often called as perspective. The camera works the same as human vision. Perspective transformation relates two different images that are alternative projections of the same three-dimensional object onto two different projective planes. This means that images taken from any viewpoint can be realized by perspective transformation, the perspective transformation can be used to generate images with different viewpoints.

We introduce perspective transformation to data augmentation and propose a perspective transformation data augmentation (PTDA) method. In this framework, perspective transformation is used to produce new data, and then "annotation alignment" is adopted to automatically generate the corresponding annotations. Fig.1 shows an overview of the proposed data augmentation for object detection. The main benefit of the PTDA method is that it is easy to implement, requires no expert knowledge, and can be applied to many different tasks. Experiments on transmission-line datasets show that PTDA method not only improves the detection performance on small dataset, but also increases the robustness of the model.

II. RELATED WORK

This section gives a brief review of the past work on data augmentation. The data augmentation techniques vary from simple transformation, learned strategies to synthesizing new images. In table 1, we give a brief summary of these data augmentation methods.

In image processing field, the widely used approaches for data augmentation include reflections, rotation and color casting [2] [22] [23], etc. Object detection methods benefit from these data augmentation techniques. Fast and Faster-RCNN use horizontal flip to augment data during train [6]] [29]. Zhong et al. [17] randomly erasing patches of the image to reduce the risk of over-fitting. Although the above techniques are available, these methods are primarily empirical and cannot be transfer to other datasets as effectively. So, recent works have focused on learning how to generate good data augmentations.

Smart augmentation is an attempt at learned data augmentation strategy, creates a network that learns how to generate augmented data [18]. AutoAugment [19] uses reinforcement learning to optimize for accuracy. More recently, Population Based Augmentation (PBA) learns non-stationary augmentation policy schedules instead of a fixed augmentation policy [24]. While the above approaches have focused on classification problems, Zoph et al. extend the previous works to object detection tasks [25]. Dvornik et al. [26] leveraged context modeling to increase the number of object instances. Zhou [30] proposed a slot-based image augmentation to generate images with more learn-able object detection related features. These learned methods are usually expensive in computation.

Another powerful data augmentation method is Generative Adversarial Nets (GAN) proposed by Goodfellow et al. [31]. The generative model in GAN framework tries to produce realistic images to fool the discriminative model, while the discriminative model attempts to distinguish the generated samples from the real images. The emergence of GAN has attracted the attention of many researchers, and many variants of GAN have been proposed to improve the quality of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

TABLE 1. Summarization of different approaches for data augmentation

	Application	Advantage	Disadvantage
Standard Augment	image classification [2] [22]; word recognition [23]; object detection [6] [17]	easy to implement	need manual selection
Learned Method	image classification [18] [19] [24]; object detection [25] [26]	automatic learning effective data augmentation policy	computationally expensive
GAN-based Method	generate image [27] [28]; image classification [20] [21]	produce images of unprecedented quality	computationally expensive; not effective for detection scenarios

synthetic image [20] [21] [27] [28]. However, the GAN approaches are rarely used for real-world scene detection tasks, because it much more difficult to generate an image with many object instances placed in a relevant background than to generate an image with only one dominant object [32]. Moreover, the generated image has no corresponding annotations.

III. PERSPECTIVE TRANSFORM FOR DATA AUGMENTATION

In this section we describe the proposed perspective transformation data augmentation (PTDA) method in details. This method takes an image and its annotation file as input, producing a new annotated output image. The mechanism of PTDA is split into two parts, shown in Fig.2. In order of computation, firstly, new images with different viewpoints were created by perspective transformation; then, annotation alignment was introduced to generate its corresponding annotation file. The combination of these two components forms a PTDA and will be described in more detail in the following sections.

A. PERSPECTIVE TRANSFORM

Assume that $\mathcal{T}_\theta$ is a perspective transformation, the pointwise transformation is:

$$\begin{aligned}\begin{pmatrix} x_s \\ y_s \\ w_s \end{pmatrix} &= \mathcal{T}_\theta\left(G_i\right) = P_\theta \begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix} \\ &= \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & 1 \end{bmatrix} \begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix},\end{aligned} \tag{1}$$

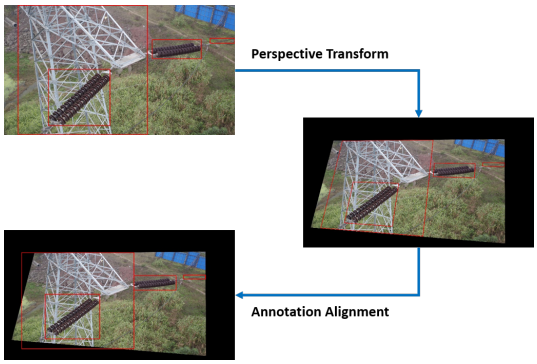


FIGURE 2. Perspective transformation data augmentation.

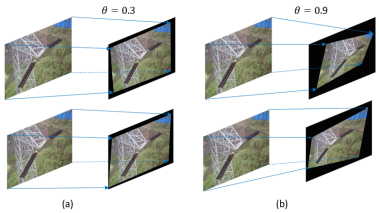


FIGURE 3. Examples of applying perspective transformation to an transmission-line image. (a) Perspective results under $\theta = 0.3$. (b) Perspective results under $\theta = 0.9$.

where $(\frac{x_s}{w_s}, \frac{y_s}{w_s})^T$ is the source coordinate of the input image pixel and $G_i = (x_t, y_t)^T$ is the target coordinate of the pixel in the output image, P_θ is the perspective transfomation matrix with 8 parameters. So, P_θ can be estimated by using four pairs of corresponding points. In estimation, we center and normalize the corresponding points. The matrix of the processed points is usually better conditioned than the matrix of the original points. On the contrary, given P_θ , the perspective transformation can be completed by the coordinates of images. To perform a perspective transform of the input image, each output pixel is calculated by multiplying with the given perspective matrix. In general, the output image has the same size as the input image, the points outside the boundary in the output image are filled with black pixels.

The use of perspective transformation for data augmentation is motivated by the perspective phenomenon in camera shooting. When the human eye views a scene, objects in the distance appear smaller than objects close by – this is known as perspective. The camera works on the same principle as human vision. The perspective transformation, which is a specific kind of planar homography, relates two different images that are alternative projections of the same three-dimensional object onto two different projective planes. This means that images taken by camera at different positions and orientation can be realized by the perspective transformation.

For simplicity, in this paper we consider the four vertices

$$A = (0, 0), B = (0, H), C = (W, H), D = (W, 0)$$

of the input image, along with the four random sampling points

$$\begin{aligned} A_\theta = (x_{tl}, y_{tl}), B_\theta &= (x_{bl}, H - y_{bl}) \\ C_\theta = (W - x_{br}, H - y_{br}), D_\theta &= (W - x_{tr}, y_{tr}) \end{aligned} \tag{2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

of the output image as corresponding points to estimate perspective matrix. Where H and W are the height and width of the image;

$$\begin{aligned} x_{tl}, x_{bl}, x_{br}, x_{tr} &\in (0, W * (\lambda\theta)), \\ y_{tl}, y_{bl}, y_{br}, y_{tr} &\in (0, H * (\lambda\theta)) \end{aligned} \tag{3}$$

are integers which represent the distance from the four points to the corresponding image boundary, they ensure any three points in A_θ, B_θ, C_θ and D_θ are non-collinear; θ is the perspective intensity parameter and we choose $\lambda = 0.3$ in the experiments of this paper. Indeed, the perspective transformation changes with the value of θ, the greater of value θ, the more obvious of perspective phenomenon. On the other hand, when the value of θ is fixed, different perspective transformations are also produced because the integers in (2) are randomly selected. The influence of θ on this transformation is shown in Fig.3. We will provide experiments on the influence of the parameter θ the section V.

B. ANNOTATION ALIGNMENT

Normally, the rectangular bounding box of the image represented by $[x_{min}, y_{min}, x_{max}, y_{max}]$, where (x_{min}, y_{min}), (x_{max}, y_{max}) are the upper-left and lower-right coordinates of the bounding box respectively. By (1), the points obtained from the four vertices of the input bounding box constitute a new bounding box. However, the resulting bounding box usually becomes trapezoidal(e.g. Fig.2), and unfortunately CNN cannot train non-rectangular bounding boxes. Therefore, we use annotation alignment to deal with the transformed bounding box.

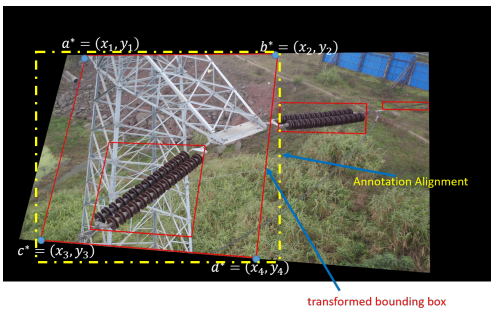


FIGURE 4. Annotation alignment. The red is the bounding boxes after perspective transformation; the yellow one is the rectangle bounding box after annotation alignment.

Assuming $a^* = (x_1, y_1)$, $b^* = (x_2, y_2)$, $c^* = (x_3, y_3)$, $d^* = (x_4, y_4)$ are the vertex coordinates of the transformed bounding box. Let us consider

$$\begin{aligned} x^*_{min} &= \min\{x_1, x_2, x_3, x_4\}, \\ x^*_{max} &= \max\{x_1, x_2, x_3, x_4\}, \\ y^*_{min} &= \min\{y_1, y_2, y_3, y_4\}, \\ y^*_{max} &= \max\{y_1, y_2, y_3, y_4\}, \end{aligned} \tag{4}$$

then $[x^*_{min}, y^*_{min}, x^*_{max}, y^*_{max}]$ is used to represent the transformed bounding box, as in Fig.4. The new bounding box can be beyond the boundaries of the actual image. The annotation alignment method makes it possible to automatically generate trainable annotated images without expensive and time-consuming manual labeling.

IV. THE TRANSMISSION-LINE IMAGE DATASETS

We now introduce the Transmission-Line Image Datasets used throughout this paper. The transmission-line images used in our study were taken by unmanned aerial vehicles (UAV). We have collected 5413 transmission-line images and constructed two original image datasets, i.e. $\mathcal{DS}_S$ and $\mathcal{DS}_L$. $\mathcal{DS}_S$ is a small dataset containing 1001 images while $\mathcal{DS}_L$ is a large dataset containing 5413 images. The images were manually annotated by labeling software (eg. LabelImage). These images are labeled with 10 categories, shown in Fig.5.

It is important to discuss the data distribution of transmission-line image Datasets. The number and size of objects in $\mathcal{DS}_S$ and $\mathcal{DS}_L$ are shown as a histogram in Fig.6. These two datasets have 17751 and 98990 annotated objects instances respectively, on average each image has 18 labels. Unlike Pascal VOC data [33], which is mainly used for the evaluation of object detection techniques, the transmission-line datasets contain a large imbalance: not only the class-wise object distribution but also the object sizes distribution. There are more than 23K wires, 13828 ceramic insulators but only 4000 composite insulators in $\mathcal{DS}_L$. Transmission-line objects with an area ration of less than 10% occupy the largest proportion of $\mathcal{DS}_L$. $\mathcal{DS}_S$ has a similar distribution as $\mathcal{DS}_L$, as shown in the second row of Fig.6.

Furthermore, during the acquisition of transmission-line images, due to the limited shooting angle of the UAV, the image at certain angles cannot be acquired. With perspective transformation, these unobtainable images can be generated from existing data.

In order to verify the validity of our proposed data augmentation algorithm, four large transmission-line datasets $\mathcal{DS}_{0.3}$, $\mathcal{DS}_{0.6}$, $\mathcal{DS}_{0.9}$ and $\mathcal{DS}_{ml}$ are constructed based on the small dataset $\mathcal{DS}_S$. The construction of these datasets is presented in Table 2. The integer in Table 2 represents the number of new images generated from each image in base-dataset under corresponding θ. For example, each image in $\mathcal{DS}_S$ generates 5 new images under $\theta = 0.3$ to form dataset $\mathcal{DS}_{0.3}$.

TABLE 2. Construction of augmented datasets.

	Base Dataset	$\theta = 0.3$	$\theta = 0.6$	$\theta = 0.9$
$\mathcal{DS}_{0.3}$	$\mathcal{DS}_S$	5	×	×
$\mathcal{DS}_{0.6}$		×	5	×
$\mathcal{DS}_{0.9}$		×	×	5
$\mathcal{DS}_{ml}$		2	2	1
VOCml	VOC2007	2	1	×

V. EXPERIMENTS AND RESULTS

In this section, we first introduce implementation details of our method. Then, we conduct extensive experiments on two datasets (the transmission-line dataset and PASCAL VOC) to verify the effectiveness of the PTDA method for object

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

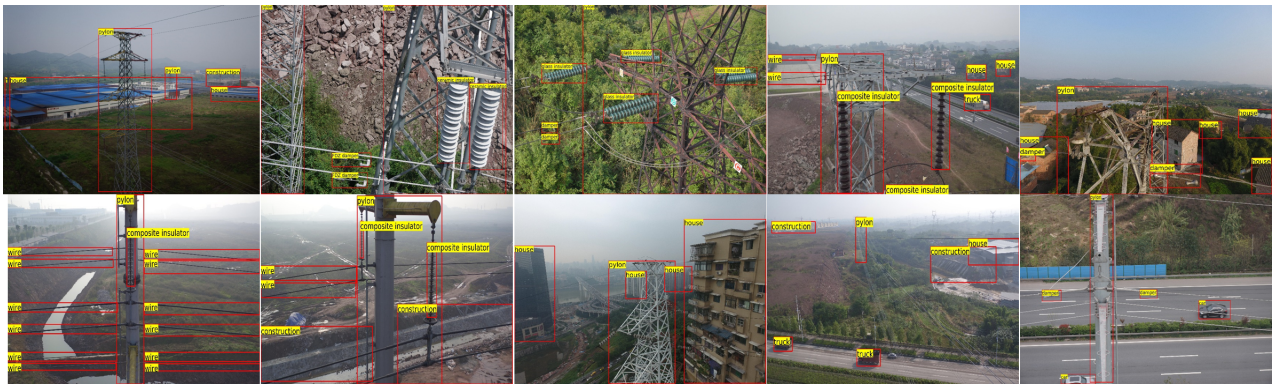


FIGURE 5. Selected examples of 10 classes in transmission-line image. The ten object classes that have been selected are: 1) Transmission-line objects: pylon, ceramic insulator, glass insulator, composite insulator, damper, wire; 2) Extended scene: construction, house, truck, car .

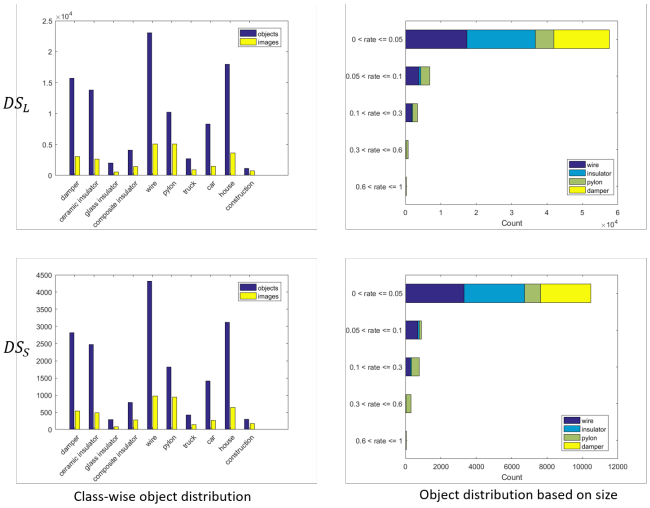


FIGURE 6. Summary of the transmission-line datasets. First column: objects and images contained in corresponding class; Second column: object distribution based on size.

detection. The hypotheses we want to test in experiments include: (1) our data augmentation technique can boost the performance in object detection; (2) by enhancing images for the classes with fewer labeled samples, it increase the learning capability for the rare classes; (3) our data augmentation technique is more effective than other data augmentation techniques.

A. IMPLEMENTATION DETAILS

We use Faster RCNN framework as the object detector. As is standard practice [6], all network backbones are pre-trained on the ImageNet1k classification set [34] and then fine-tuned on the detection dataset. We use the publicly available pre-trained VGG-16 model [3]. The detector is trained with SGD+Momentum. We use a weight decay of 0.0005 and a momentum of 0.9. The learning rate is 10^{-3} initially and we decay it by 0.1 every 50K steps.

Each worker takes a single input image per step, the batch size for RPN and box classifier training are 64 and 256 respectively. Input images are resized to have 600 minimum pixels and 1000 maximum pixels while maintaining the aspect ratio. In [6], during the training of Faster RCNN, the images were horizontally flipped with a probability of 0.5% for data augmentation. In order to compare with other data augmentation methods, no horizontally flipping is used during our training.

B. EXPERIMENTS ON TRANSMISSION-LINE DATASETS

The first step in understanding the benefits of perspective transformation data augmentation is to compare the detectors trained on augmented datasets with the detectors trained on original datasets. We comprehensively evaluated the PTDA method on the six transmission-line datasets mentioned earlier, namely $\mathcal{DS}_S$, $\mathcal{DS}_L$, $\mathcal{DS}_{0.3}$, $\mathcal{DS}_{0.6}$, $\mathcal{DS}_{0.9}$ and $\mathcal{DS}_{ml}$. All these datasets were divided into three parts in the same proportion, i.e. $\frac{3}{5}$ for training, $\frac{1}{5}$ for validation and $\frac{1}{5}$ for testing. The detectors were trained on the train sets and evaluated on its corresponding validation sets. We report the final results on the $\mathcal{DS}_L-$test and $\mathcal{DS}_{ml}-$test. We use mean average precision (mAP) to measure the accuracy of object detectors.

1) $\mathcal{DS}_L-$test results

Table 3 shows the results on $\mathcal{DS}_L-$test. Obviously, detectors trained on large datasets perform better than those trained on small datasets. Our PTDA method provides a significant improvement in object detection accuracy. When training on augmented datasets, the detection results are up to 17.5% better than detector trained on small original dataset $\mathcal{DS}_S$ in mAP. Moreover, when we trained detector on $\mathcal{DS}_{0.3}$, which has similar amount of data as the large original dataset $\mathcal{DS}_L$, it is even more accurate, surpassing $\mathcal{DS}_L$ by 6.2% mAP. It indicates that our augmented dataset with $\theta = 0.3$ yields superior results with respect to the similar amount original image dataset.

2) $\mathcal{DS}_{ml}-$test results

$\mathcal{DS}_{ml}-$test is composed of augmented images, its results are shown in Table 4. We can see the similar performance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

TABLE 3. $\mathcal{DS}_L$-test detection average precision (%) using different training data

train set	mAP	ceramic insulator	glass insulator	composite insulator	damper	pylon	wire	truck	car	house	construction
$\mathcal{DS}_S$	40.8	53.4	51.8	41.5	52.0	68.2	43.1	13.3	14.5	39.1	31.2
$\mathcal{DS}_L$	52.1	62.4	63.5	44.1	54.5	78.9	62.4	25.8	21.6	66.2	42.3
$\mathcal{DS}_{0.3}$	**58.3**	63.0	63.5	35.3	63.3	80.6	76.3	25.5	24.0	71.8	80.1
$\mathcal{DS}_{0.6}$	51.6	53.8	63.5	35.7	54.6	71.0	62.4	19.8	15.7	67.9	72.0
$\mathcal{DS}_{0.9}$	48.9	53.8	63.5	35.0	53.1	70.1	53.8	16.2	16.7	64.2	62.3
$\mathcal{DS}_{ml}$	55.3	62.5	63.6	35.5	54.2	79.4	68.1	23.5	17.6	70.3	78.1

TABLE 4. $\mathcal{DS}_{ml}$-test detection average precision (%) using different training data

train set	mAP	ceramic Insulator	glass Insulator	composite Insulator	damper	pylon	wire	truck	car	house	construction
$\mathcal{DS}_S$	39.5	44.3	53.9	33.7	26.7	62.2	47.5	17.8	14.0	42.6	51.9
$\mathcal{DS}_L$	39.8	44.2	53.5	34.3	34.9	68.1	54.7	17.2	13.0	47.1	31.5
$\mathcal{DS}_{0.3}$	**51.9**	52.4	63.0	42.6	35.0	71.8	70.0	24.7	16.7	62.6	79.9
$\mathcal{DS}_{0.6}$	50.3	45.1	62.5	35.7	34.4	72.0	75.5	20.4	15.9	62.1	80.0
$\mathcal{DS}_{0.9}$	48.8	45.0	63.2	35.8	31.1	72.0	68.8	17.2	15.0	61.4	78.7
$\mathcal{DS}_{ml}$	50.5	45.0	62.7	35.5	35.1	72.0	75.2	21.3	16.7	62.1	79.0

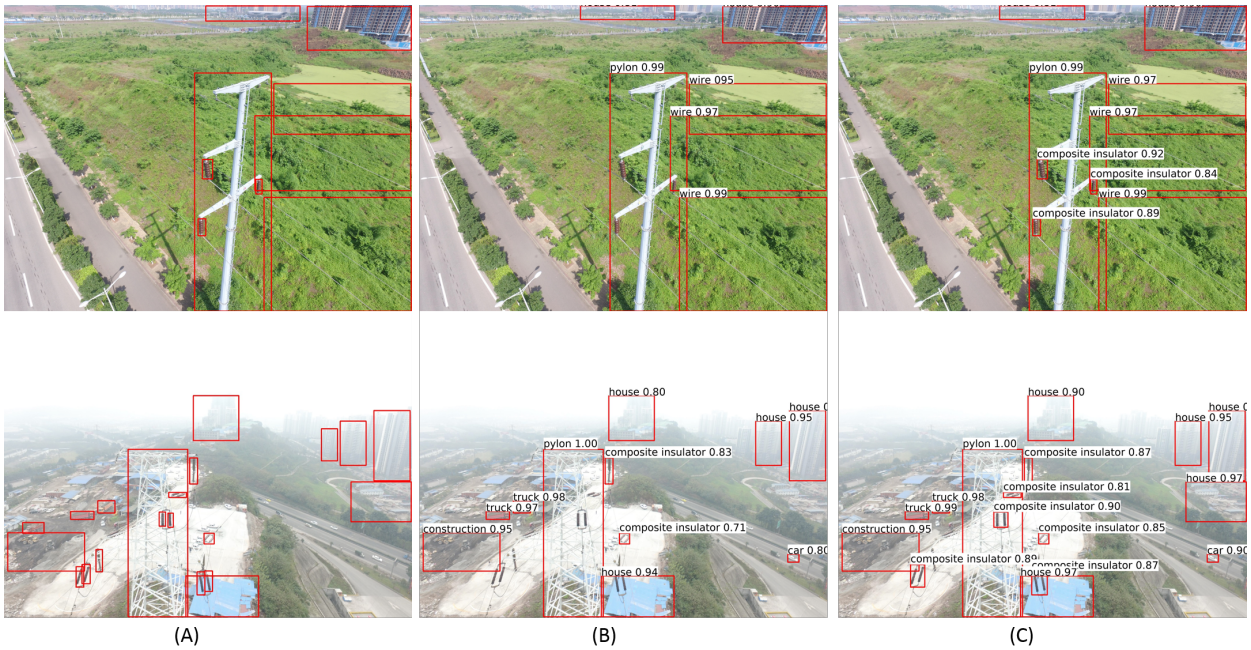


FIGURE 7. Comparison of detection results. (A)the original image with ground truth bounding box; (B)the detection results trained by $C\mathcal{DS}_{ml}$; (C)the detection results trained by $C\mathcal{DS}_{bl}$.

trend as we observed on $\mathcal{DS}_L$-test. Detector trained on $\mathcal{DS}_{0.3}$ outperforms both small and same amount original image dataset. Interestingly, we observe that the performance of $\mathcal{DS}_L$ is only 0.3% better than $\mathcal{DS}_S$ in mAP, and far below the improvement of augmented datasets. This shows that the PTDA method makes detectors have better generalization performance. Because perspective transform can fabricate new images taken by cameras at different viewpoints, augmented datasets have better object diversity than original datasets.

3) Impact of Hyper-Parameter

When implementing PTDA method for CNN training, we need to evaluate one hyper-parameter θ. To demonstrate the impact of this hyper-parameter on the model performance, we compare the results of the four augmented datasets.

Table 3 and Table 4 show that, the detector trained by $\mathcal{DS}_{0.3}$ achieves the top results on both $\mathcal{DS}_L$-test (58.3%) and $\mathcal{DS}_{ml}$-test (51.9%). The performance of $\mathcal{DS}_{ml}$ is the second on both test data. Although all the augmented datasets have better performance than the original datasets on $\mathcal{DS}_{ml}$-test, $\mathcal{DS}_{0.6}$ and $\mathcal{DS}_{0.9}$ are less effective than $\mathcal{DS}_L$ on $\mathcal{DS}_L - test$. This is not surprising, because perspective transformation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

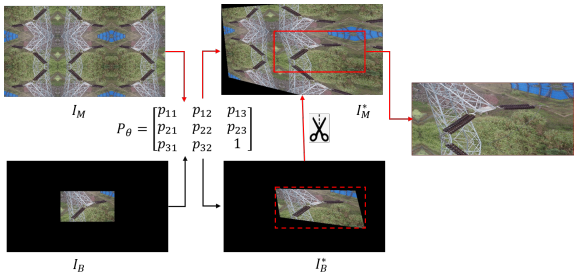


FIGURE 8. Ppadding black edges with "Crop".

TABLE 5. The performance of padding black edges(%)

test set \ train set	$\mathcal{RDS}_{0.3}$	$\mathcal{RDS}_{0.6}$	$\mathcal{RDS}_{0.9}$	$\mathcal{CDS}_{ml}$
$\mathcal{DS}_L$-test	**62.8**	51.7	43.5	**62.5**

TABLE 6. Average precision(AP) of ceramic insulators(%)

train set	$\mathcal{RDS}_{0.3}$	$\mathcal{CDS}_{ml}$	$\mathcal{CDS}_{bl}$
AP	52.3	53.0	**61.2**

increases the diversity of objects while producing black edges(Fig.3). With the increase of θ, the black area of the image is also increasing. When θ is too large, the black edge seriously affects the quality of image. In the next experiments, we will discuss the influence of padding black edges.

4) Impact of Padding Black Edges

In section II, we fill the perspective transformed image with black pixels. In this experiment, we evaluate two approaches for padding black edges with the mirror of original image.

The first method ("Reflect") directly reflects the image content along the boundary to fill the black border. The second method ("Crop") cuts a non-black-edge image in the transformed mirror image. Fig.8 shows the overview of "Crop". For "Crop", we first flip the original image along its four sides to form a mirror image $\boldsymbol{I}_M$. Similarly, fill the surrounding area of the original image with black pixels to form image $\boldsymbol{I}_B$ of the same size as $\boldsymbol{I}_M$. Then, a same perspective transformation is performed on $\boldsymbol{I}_M$ and $\boldsymbol{I}_B$, obtaining $\boldsymbol{I}_M^*$ and $\boldsymbol{I}_B^*$. Finally, we find the smallest rectangle containing image area in $\boldsymbol{I}_B^*$ and crop the same area in $\boldsymbol{I}_M^*$. The cropped area is the padded image we need.

For these two methods, the padded part may contain parts of the objects, but for simplicity we ignore these objects when training a detector. The image padded by "Reflect" has the same size as the original image, while the image padded by "Crop" is smaller than the original image. The biggest difference between "Reflect" and "Crop" is that the padding area in "Reflect" grows larger with the increases of θ, but the padding area is insensitive to θ in "Crop". So in the experiments, only $\mathcal{DS}_{ml}$ is padded by "Crop", generating $\mathcal{CDS}_{ml}$; $\mathcal{DS}_{0.3}$, $\mathcal{DS}_{0.6}$ and $\mathcal{DS}_{0.9}$ are padded by "Reflect", generating $\mathcal{RDS}_{0.3}$, $\mathcal{RDS}_{0.6}$ and $\mathcal{RDS}_{0.9}$. Table 5 presents the results on $\mathcal{DS}_L$-test.

We observed that, on one hand, detectors trained on $\mathcal{RDS}_{0.3}$ and $\mathcal{CDS}_{ml}$ outperform the unfilled datasets by a large margin. Furthermore, $\mathcal{CDS}_{ml}$ achieves approximately equal performance to $\mathcal{RDS}_{0.3}$. This illustrates that padding black edges with image information reduces the effect of noise. On the other hand, applying padding on $\mathcal{RDS}_{0.6}$ and $\mathcal{RDS}_{0.9}$ fails to improve the accuracy. This is because, as the equilibrium parameter θ increases, the number of ignored objects in padded area increases quickly in "Reflect". So if not specified, we use "Crop" to fill the black edges in the following experiment.

C. DATA AUGMENTATION FOR UNBALANCED DATA

The data distribution of transmission-line datasets is very uneven, take composite insulator as an example. All the images we collected contain 3935 composite insulators, only 4% of the total annotated objects, compared with 14% for ceramic insulator, $23\%$for wire. In particular, there are even less composite insulators in the form shown in Fig.7. This results in low detection accuracy for this type of composite insulator, even though data augmentation and padding are used (Table 6). Fig.7 column (B) shows a detection result on an image using the detector trained on $\mathcal{CDS}_{ml}$, there are still some composite insulators cannot be detected.

We augment specific category to address data imbalance issues. In this experiment, we augment the composite insulator samples through our PTDA strategy. In $\mathcal{DS}_L$, if an image contains ceramic insulator of the form shown in Fig.7, this image generates 5 new images under multiple equilibrium parameter θ values, similar to $\mathcal{DS}_{ml}$. Other composite insulator-containing images generate one new image under $\theta = 0.9$. All the generated images are padding black edges by "Crop". Then the augmented images and $\mathcal{DS}_L$ are combined into a new dataset $\mathcal{CDS}_{bl}$.

Table 6 shows the average precision(AP) of ceramic insulators on $\mathcal{DS}_L - test$. It's clearly that the insulator augmented dataset $\mathcal{CDS}_{bl}$ achieves the best performance(61.2% in AP) on composite insulators compared to $\mathcal{RDS}_{0.3}$ and $\mathcal{CDS}_{ml}$. Fig.7 column (C) shows the detection results of the detector trained by $\mathcal{CDS}_{bl}$, and we have effectively detected the composite insulators that have not been detected before. So, the PTDA can not only augments the whole dataset but also can used to generate objects in rare classes to increase the learning capability for the rare classes.

D. EXPERIMENTS ON PASCAL VOC

We further verify the validity of our PTDA method on Pascal VOC dataset. VOC2007 is used to generate a new 3x dataset VOCml(Table 2). Padding black edges is used for the images in VOCml. VOCml is split into 50% for training/validation and 50% for testing, the same as VOC2007. The detector is trained with three training sets, VOCml trainval, VOC2007

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962572, IEEE Access

TABLE 7. Object detection mAP (%) on the PASCAL VOC2007 test sets.

training data	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	TV
VOC07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
VOC07&12 [6]	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
VOCml	**79.9**	85.1	86.6	78.6	75.7	65.2	83.5	88.4	88.9	65.8	83.6	74.3	86.4	84.7	85.5	88.0	62.0	75.5	75.3	87.7	76.3

TABLE 8. The detection average precision of different data augmentation methods(%).

method	mAP	
	$\mathcal{DS}_L-$test	VOC2007 test
PTDA + "Crop"	**62.5**	**79.9**
Context-DA [26]	59.4	75.9
lot-based image augmentation [30]	60.7	73.1
Flipping [6]	47.6	69.9
Adding Noise	55.1	75.8
Random Erasing [17]	51.9	76.2

trainval and VOC2007&2012 trainval. In particular, during the training on VOC2007 trainval and VOC2007&2012 trainval, images are horizontally flipped at the probability of 0.5.

Table 7 shows the object detection results on PASCAL VOC2007-test. When training with VOC2007 trainval, the baseline is 69.1% mAP. The detector trained with our augmented dataset achieves an improvement to 79.9% mAP, obtaining a 10.8% increase. When using the VOC07&12 trainval set, the baseline is 73.2% which is better than only using VOC07 trainval set. Again, VOCml outperforms the VOC07&12 trainval baseline significantly, by 6.7%.

E. COMPARISON TO OTHER METHODS

To evaluate the usefulness of different data augmentation techniques, we trained new detectors on datasets that augmented with different data augmentation methods. In this experiment, we compare our PTDA with several common algorithms, including: adding random noise, affine transformation (rotation, flipping etc.) [6], random erasing [17]. In addition, we also compare with the two learned methods, Context-DA [26] and slot-based image augmentation [30], to illustrate the benefits of our technique.

These methods are used to augment datasets $\mathcal{DS}_S$ and VOC2007. In PTDA method, we use "Crop" to fill the black edges. The datasets obtained by PTDA + "Crop" are $C\mathcal{DS}_{ml}$ and VOCml mentioned above. In "Flipping" , the images in base datasets $\mathcal{DS}_S$ and VOC2007 are horizontally flipped with probability 0.5, the same as in [6]. Similarly,the original images are added with different levels of gaussian noise in "Adding Noise". All the configurations of other methods remain the same as the original implementation. In the transmission-line object detection task, the augmentated datasets are the same size as $C\mathcal{DS}_{ml}$ and distributed in the form of train:val:tes = $3:1:1$. In the VOC object detection task, the size of the augmentated datasets are the same as the VOCml and these datasets are distributed as train:val:tes = $1:1:2$.

The detection average precision of these data augmentation methods on $\mathcal{DS}_L-$test and VOC2007-test are presented in Table 8. It is clearly that, our method is superior to standard data augmentation methods and slightly better than the learned methods in both transmission-line dataset and PASCAL VOC dataset.

F. DISCUSSION

A summary of the performance of our PTDA approach is shown in Table$3-8$. Table$3-5$ show the performance of our method on transmission-line dataset. The results have preliminarily shown the benefits of the PTDA strategy, with improvements of ranging from $+3.2$ mAP to $+17.5$ mAP. When implementing "Crop" to pad the black edges, the PTDA + "Crop" achieves a further increase by $+4.2$ mAP. Considering the important but rare categories, we tried to augmentation the specific category. Table 6 indicates that our method can be applied to rare categories to improve their detection accuracy. Additionally, we extended our method to PASCOL VOC detection task and lead to a substantial improvement. The results in Table 7 show that the PTDA policy transfers well across different tasks. Table 8 compares the detection results of detectors trained with different data augmentation method. As can be seen, our proposed method yields the best results on both transmission-line dataset and PASCAL VOC dataset, and provides $+10.4$ mAP and $+10.0$ mAP improvement respectively.

VI. CONCLUTION

In this paper, we introduce a data augmentation technique for object detection based on perspective transformation. The proposed PTDA technique automatically generates new data with corresponding annotations without expert knowledge. According to our experimental results, PTDA has great advantages, which not only improves the detection performance on small datasets, but also the robustness of the detection model. In the future work, we will further strengthening the detection for small objects. In addition, since there is a hyperparameter θ in our method affects the performance of the model, we also plan to learn an effective θ for a target dataset.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in The IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in The IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.
[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in The IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.
[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European conference on computer vision. Springer, 2016, pp. 20–36.
[12] A. Diba, V. Sharma, and L. Van Gool, "Deep temporal linear encoding networks," in The IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2329–2338.
[13] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in The IEEE international conference on computer vision, 2017, pp. 843–852.
[14] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," in International Conference on Learning Representations, 2016.
[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.
[16] D. Han, Q. Liu, and W. Fan, "A new image classification method using cnn transfer learning and web data augmentation," Expert Systems with Applications, vol. 95, pp. 43–56, 2018.
[17] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," arXiv preprint arXiv:1708.04896, 2017.
[18] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," IEEE Access, vol. 5, pp. 5858–5869, 2017.
[19] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," arXiv preprint arXiv:1805.09501, 2018.
[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.
[21] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," in International Conference on Learning Representations, 2018.
[22] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," arXiv preprint arXiv:1501.02876, 2015.
[23] L. S. Yaeger, R. F. Lyon, and B. J. Webb, "Effective training of a neural network character classifier for word recognition," in Advances in neural information processing systems, 1997, pp. 807–816.
[24] D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in International Conference on Machine Learning, 2019, pp. 2731–2741.
[25] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," arXiv preprint arXiv:1906.11172, 2019.
[26] N. Dvornik, J. Mairal, and C. Schmid, "Modeling visual context is key to augmenting object detection datasets," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 364–380.
[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in International Conference on Learning Representations, 2018.
[28] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks. arxiv," in International Conference on Learning Representations, 2017.
[29] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[30] Y. Zhou, "Slot based image augmentation system for object detection," arXiv preprint arXiv:1907.12900, 2019.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.
[32] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," arXiv preprint arXiv:1809.03193, 2018.
[33] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," International journal of computer vision, vol. 111, no. 1, pp. 98–136, 2015.
[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," International journal of computer vision, vol. 115, no. 3, pp. 211–252, 2015.

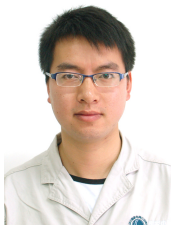
KE WANG received the B.S. degree in school of mathematics and statistics from Xinyang Normal University, Xinyang, China, the M.S. degree in college of mathematics and statistics from Chongqing University, Chongqing, China. She is currently pursuing the Ph.D. degree in college of computer science at Chongqing University, Chongqing, China. Her research interests include computer vision, machine learning, and pattern recognition.

BIN FANG received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, the M.S. degree in electrical engineering from Sichuan University, Chengdu, China, and the Ph.D. degree in electrical engineering from the University of Hong Kong, Hong Kong. He is currently a Professor with the Department of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, medical image processing, biometrics applications, and document analysis. He has published more than 100 technical papers and is an Associate Editor of the International Journal of Pattern Recognition and Artificial Intelligence.

JIYE QIAN received the B.S. degree in computer science from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2004. He received the M.S. degree in automation and Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2007 and 2012, respectively. He is currently working on joint postdoctoral program with Chongqing University and State Grid Chongqing Electric Power Company, Chongqing, China. His research interests include computer vision, machine learning, and sensor signal processing.

SU YANG received the B.E. degree in computer science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2016. He is currently studying in Chongqing University, Chongqing, China, to pursue his master's degree. His research interests include computer vision, machine learning.

XIN ZHOU received the B.E. degree in the Internet of Things Engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2017. He is currently studying in Chongqing University, Chongqing, China, to pursue his master's degree. He is majoring in computer science, whose research interests include pattern recognition, computer vision, machine learning and deep learning.



JIE ZHOU acquired a bachelor degree in Electrical Engineering and Automation, Chongqing University in 2012. He currently works in the State Grid Chongqing Yongchuan Power Supply Company. His research interest includes operation and maintenance of transmission line,UAV patrol inspection,artificial intelligence and etc.