Multi-Scale Feature Enhanced Domain Adaptive Object Detection For Power Transmission Line Inspection

PENGYU ZHANG1, (Student Member, IEEE), ZHE ZHANG2, YANPENG HAO3, (Member, IEEE), ZHIHENG ZHOU1, (Member, IEEE), BING LUO4,5, AND TINGTING WANG4,5

1School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China
2Pazhou Lab, Guangzhou 510335, China
3School of Electronic Power Engineering, South China University of Technology, Guangzhou 510641, China
4CSG Electronic Power Research Institute Co.Ltd, Guangzhou 510080, China
5Electrical Power Research Institute of China South Grid, Guangzhou 510080, China

Corresponding author: Zhe Zhang (zhangz@gcu.edu.cn)

This work was supported by the open fund of National Engineering Laboratory for Ultra High Voltage Engineering Technology (Kunming, Guangzhou) (Grant no. NEL201807).

ABSTRACT Domain adaptive object detection aims to build an object detector for the unlabeled target domain by transferring knowledge from a well-labeled source domain, which can alleviate the problem of cumbersome labeling of object detection in cross-scene power transmission line inspection. Remarkable advances are made recently by mitigating distributional shifts via hierarchical domain feature alignment training of detection networks. However, domain adaptive object detection is still limited in learning the invariance representation of multi-scale features. Specifically, the scale of objects varies in the scenes of aerial inspection, which hinders the knowledge transfer from the labeled source domain. In this paper, we propose a multi-scale feature enhanced domain adaptation method for cross-domain object detection of power transmission lines inspection. The proposed method consists of two components: 1) Multi-Scale Fusion Feature Alignment module (MSFA) to strengthen similar representation characteristics of different scales object in domain adaptive by utilizing context information conveyed from other levels; 2) Multi-Scale Consistency Regularization module (MSCR) to jointly optimize the multi-scale feature learning of each level, which promotes domain invariant feature learning at each level. Experimental results demonstrate that our method significantly increases the performance of the object detector in several cross-scene transmission line inspection tasks.

INDEX TERMS Domain adaptation, object detection, deep learning, power transmission lines inspection, fault detection, insulator.

I. INTRODUCTION

Power transmission line protection [1]–[3] is an essential issue in power system engineering. The power transmission line components, such as insulators, shockproof, clamps, may wear, be teared, or suffer other forms of damage, which will not only affect electricity delivery but also cause more dangerous consequences. Therefore, regular inspection and monitoring of the key components in the transmission line is a crucial scheme to ensure the safety of the power system.

In recent years, unmanned aerial vehicles (UAV) instead of the traditional inspection way have become the common tool for intelligent requirements in transmission lines inspection [4], [5]. Hence, obtain a high precision object detector from aerial images is the critical technology in transmission line inspection for intelligent requirements [6], [7]. With the development of hardware equipment and deep learning theory, object detection has made a significant breakthrough in the task of automated circuit inspection [7]–[10]. However, such deep models usually need a large-scale annotated dataset for supervised learning. Unfortunately, the cost of collecting and annotations is usually expensive:
for an image, researchers must annotate each instance with categories, coordinates and precise bounding boxes. Moreover, it is not always feasible to collect sufficient annotated training samples from the new environments. For example, UAV is expected to perform inspection tasks reliably in various weather situations (such as rainy days), but the training samples are usually collected in normal weather with clear vision. In the real world, scenes, weather, lighting conditions, and camera settings will cause domain differences, for example, the appearance of objects, background and image quality, etc. Such domain discrepancy or domain-shift can cause the model to perform poorly in a new domain [11].

In such situations, unsupervised domain adaptation (UDA) is an appealing solution to alleviate cumbersome labeling for the unlabeled target domain by transferring knowledge from a well-labeled source domain [12]. In UDA, general practice is to align target with the source distribution, so that the model can generalize to the target domain without annotation. Several works have recently attempted to apply adversarial learning of UDA to the cross-domain adaptive object detection research. Current methods typically minimize the domain discrepancy at multiple levels [13]–[16]. Chen et al. [17] reduce the domain discrepancy on both image and instance level. Furthermore, [18] integrates an image-level multi-label classifier upon the domain adaptive faster r-cnn [17] as categorical regularization to match crucial image regions and important instances across domains.

Similarly, He et al. [14] propose a hierarchical domain feature alignment module to align the multi-level features. Furthermore, [14] adopts a weighted gradient reversal layer to re-weight training samples. Zhuang et al. [16] propose a multi-level features alignment framework which trains adversarial domain classifiers in a hierarchically-nested fashion. They hold a common belief that multi-level adversarial adaptation helps the feature learning of domain invariant.

In deep convolutional networks, each layer’s receptive field is fixed. Feature maps of the shallow layer have small receptive fields that are responsible for small objects, while semantic-rich features in deep layers have large receptive fields and are more fitting for large objects [19], [20]. The previous multi-level alignment method adopts independently align the features on each layer to decrease the cross-domain discrepancy, however, they did not do the object’s features of different scales maintain similar features representation which may impair domain invariant representation and lead to worse detection results in the target domain. As shown in figure 1, the instance scale of equipment in the inspection image between domains is very different. For example, the scale of the PolymerInsulator in the source image is much larger than the one in the target image, which hinders the knowledge transfer from the labeled source domain. Some research shows that shallow and deep features are complementary to each other, and their combinations are beneficial for yielding the feature maps with similar
II. RELATED WORK

In this section, we briefly review several study topics relevant to our work, including object detection and domain adaptation.

Object Detection. As a basic task in computer vision, object detection aims to locate and classify all instances of different objects within images. Sliding window algorithms with hand-crafted features [23]–[25] are first used for object detection. A more powerful backbone feature extractor helps model to generate detection results more accurately. Due to the development of deep convolutional neural networks, the recently advanced object detection methods [26], [27] based on CNN have achieved excellent results. Based on the network forwarding pipelines, these methods can be categorized into one-stage and two-stage detectors. RCNN [28] is one of the first studies that employ two-stage detectors; it first generates a sparse set of proposals by selective search, and then classifies the proposed region with a region classifiers. After RCNN, Faster R-CNN [29] introduces a region proposal network (RPN) to predict object proposals, which is trained on the feature map with supervision. Mask R-CNN [30] extends Faster R-CNN through combining a segmentation masks component, utilizing multi-task learning to improve the detection performance. Cascade R-CNN [31] designs a multi-stage detector base on Faster R-CNN which refine proposals in a cascaded framework. SSD [32] exploits a one-stage paradigm, merging all the predicted category and bounding box from multiple feature maps. YOLO [33] regards object detection as a regression task, where each sample is divided into set number grid cells on spatially. The detector predicts the bounding box and the category of the object with the centers of objects at each cell. Recently, anchor-free is popularized, CornetNet [34] introduces a novel anchor-free framework that the bounding boxes can be detected as a pair of corners. Similarly, [35] proposes a novel anchor-free, two-stage framework to extract corner keypoint combinations and applies two-step classification to filter out false positives. Considering robustness and flexibility, we employ Faster R-CNN [29] as the backbone of our framework.

Domain Adaptation. In recent years, many domain adaptation algorithms have been proposed, attempting to decrease annotation cost by learning domain invariant features (shared features) between domains [36]–[40]. The critical technical challenge of domain adaptation is how to reduce the feature distribution discrepancy between different domains. Theoretical analyses [41] have shown that the target risk of a model can be minimized by limiting the risk on the source samples and the distribution discrepancies between the source and target domains. As a result, existing domain adaptation methods mainly focus on minimizing the statistical distance [42], e.g., maximum mean discrepancy (MMD) [43]–[46], second-order statistics (covariances) [47]–[49], geodesic distance [50], [51]. In [52]–[54] a reproducing kernel Hilbert embedding the hidden features in the network is learned and mean-embedding matching is performed for both domain distributions. Inspired by the generative adversarial network [55], adversarial learning is successfully applied to feature distribution alignment of different domains. In [56], [57], the adversarial algorithm and domain classifier are trained together to extract shared features that are discriminable and domain invariant. Reference [58] proposes a class-conditioned domain alignment method to reduce domain class imbalance and cross-domain class distribution shift. Some recent UDA methods [59], [60] utilize MIXUP [61] to regularization the domain classifier, which encourage learn domain-invariant features representations across domains. Much research has achieved compelling performance on image classification [36]–[40], [56], [57], while some study on the more complicated task such as semantic segmentation [62]–[65]. Object detection task is more complicated than them, therefore, how to build a detector for the unlabeled target domain is indeed a challenging issue.

III. METHODOLOGY

In this section, we first describe the formal definitions for the problem and then introduce the proposed multi-scale fusion feature alignment that utilizes context fusion and feature
alignment. To improve precision and robustness, we further combine multi-scale consistency regularization that takes consistency into account during adaptation. In the domain adaptive detection problem, we are given a source domain $D_s = \{(x_s^i, b_s^i, y_s^i)\}$ and a target domain $D_t = \{x_t^i\}$. Here, $x_s^i$ is an image, $b_s^i$ is the corresponding bounding box and $y_s^i$ is the foreground class label ($y_s^i \in \{1, 2, 3, \cdots, K\}$) of $x_s^i$. We set the class label to 0 for the backgrounds. The source domain $D_s$ and the target domain $D_t$ have the same label space but follow a different distribution. Due to the distribution shift, a detection model trained on $D_s$ cannot generate accurate results on $D_t$. To solve this problem, current studies [13]–[16] follow the theoretical analysis [41] to minimize the domain discrepancy at multiple levels. However, they are still limited in learning the invariance representation of multi-scale features. Hence, in this work, we design a hierarchical domain adaptation network to construct a multi-scale fusion feature and learn more similar representation characteristics of different scales objects. In this way, the discrepancy of multi-scale fusion feature between source and target data can be mitigated and the detector trained on labeled source data can be transferred to the target domain effectively. The architecture of our proposed multi-scale feature enhanced domain adaptive detection model is illustrated in Figure 2. We aim to obtain high discriminative feature representation at each scale alignment, hence exploiting the multi-level context information as much as possible. In the following, details of the proposed network will be explained.

A. MULTI-SCALE FUSION FEATURE ALIGNMENT

Unlike former approaches that independently align different level features, our method is to strengthen the similar representation characteristics of different scales object in domain adaptive by utilizing context information conveyed from other levels. The architecture is shown in Figure 2. It includes two steps, which are context fusion and feature alignment.

**Context Fusion.** High-level features have more semantic meanings in backbone networks, and the low-level features have spatial-rich knowledge. This inspires us that the low-level and high-level information is complementary for domain adaptive object detection. To explore multi-scale features and preserve their semantic hierarchy simultaneously, we integrate context information conveyed from other levels.
We used the VGG network [66] as the backbone and aligned the features of block 3, block 4, and block 5. Features of block $k$ are denoted as $f_k$. Take the features $f_4$ of block 4 as an example. We first resize features of other scale (features $f_3$ of block 3 and features $f_5$ of block 5) to the same scale of block 4, with downsampling or upsampling. Once the features are rescaled, the context features of block 4 are obtained by simple averaging as

$$f_{4}^{\text{context}} = f_{\text{rescaled}}^{3} + f_{\text{rescaled}}^{5}$$  \hfill (1)

The context fusion module (CFM) in Figure 2 shows the process of acquiring the context features of block 4. We get the context fusion features of block 4 by concatenating the original features and the corresponding context features

$$f_{4}^{\prime} = [f_{4}, f_{4}^{\text{context}}].$$  \hfill (2)

In the same way, we got the context fusion features $f_{3}^{\prime}$ and $f_{5}^{\prime}$ from block 3 and block 5, respectively.

**Feature Alignment.** The critical technical challenge of domain adaptive detection is to decrease the distribution discrepancy across the two domains. To achieve this goal, we apply the domain adversarial training to construct a two-player minimax game. The first player is multi-scale domain classifiers $G_{d}^{k}$ that integrates into blocks 3, 4, 5 in the backbone network, it tries to distinguish the multi-scale fusion feature representations between the source and target domains, while the second player is the feature learning network $F$, it aims to confuse the domain classifiers $G_{d}^{k}$. We notice that combinations of the context information are beneficial for the multi-scale discriminative feature learning. Hence, we constructed three-domain classifiers of different scales for context fusion features $f_{i}^{k}$ to alignment the source and target domains. The domain label of $f_{i}^{k}$ is set to 1 if the image is sampled from the source domain, and the domain label of $f_{i}^{k}$ is set to 0 if the image is sampled from the target domain. Then the loss function of $k^{th}$ domain classifiers $G_{d}^{k}$ can be formulated as

$$L_{G_{d}^{k}}(G_{d}^{k}, F) = E_{j \sim D_{s}} \log \left( G_{d}^{k}(\hat{f}_{j}^{k}) \right) + E_{j \sim D_{t}} \log \left( 1 - G_{d}^{k}(\hat{f}_{j}^{k}) \right)$$  \hfill (3)

where the domain classifiers $G_{d}^{k}$ is trained to distinguish the feature of the source and target domains, while the feature learning network $F$ is finetuned simultaneously to confuse the domain classifiers. We use the Gradient Reverse Layer (GRL) [67] to implement the domain adaptive training, the sign of gradients is reversed when training a feature extract network.

**B. MULTI-SCALE CONSISTENCY REGULARIZATION**

Note that the multi-scale fusion feature alignment promoted the semantic alignment of images at each level since it can obtain features with more similarly semantic representation across the scales. Although different scales of domain classifiers are applied to learning multi-scale domain invariant features, each sub-network path tries to minimax the domain classifier loss independently. Yosinski et al. [68] show that the features must eventually transition among domain-agnostic and domain-specific by the network. Hence, the domain invariant features obtained from different layers may be different among domains. Thus, we design the multi-scale consistency regularization module to make the multi-scale domain invariant feature learning consistent among different layers.

First of all, we perform scale consistency transformation on the fusion feature maps of each scale. To increase the training efficiency of the adversarial domain classifiers, we down-scale the fusion feature maps $f_{3}^{\prime}$ and $f_{4}^{\prime}$ to the same size of $f_{5}^{\prime}$. To decrease the size of a fusion feature map without losing context enhanced information, we used the Space To Depth (STD) method to implement down-scale [69]. Figure 3 shows an intuitive example of STD. Context enhanced features at block 3 and 4 after processing from scale consistency model (SCM) are denoted as $\hat{g}_{3}$ and $\hat{g}_{4}$, respectively. For convenience, $f_{5}^{\prime}$ is denoted as $\hat{g}_{5}$. Secondly, we construct multiple domain classifiers with the same scale and do consistency constraints on multiple domain classifiers. The consistency regularizer loss can be written as

$$\mathcal{L}_{\text{con}}(\hat{g}_{d}^{3}, \hat{g}_{d}^{4}, \hat{g}_{d}^{5}) = \sum_{i,j=3,4,5} \left\| \hat{g}_{d}^{i}(\hat{g}_{j}) - \hat{g}_{d}^{i}(\hat{g}_{j}) \right\|^{2}_{2}$$  \hfill (4)

![FIGURE 3. Intuitive examples of space to depth. During the space to depth method (top), we aim to decrease the size of the feature map without losing information. (a) is an example of a feature map before the transformation. (b) is an example of the transformed feature maps.](image-url)
C. NETWORK OPTIMIZATION

With the derivation above, the total optimization of adversarial alignment and detection model can be formulated as

\[
\mathcal{L}_{\text{dec}}(F) + \lambda \mathcal{L}_{\text{con}}(\tilde{G}_d^1, \tilde{G}_d^4, \tilde{G}_d^5) + \lambda \sum_{k=3}^{5} \left( \mathcal{L}_{G_d^k}(G_d^k, F) + \mathcal{L}_{\tilde{G}_d^k}(\tilde{G}_d^k) \right)
\]  

(6)

where \( \mathcal{L}_{\text{dec}}(F) \) noted the detection loss of Faster-RCNN, which contains classification loss and regression loss of RPN and heads, and \( \lambda \) is a hyper-parameter used to balance the detection loss and domain adaptation loss.

IV. EXPERIMENTS

A. DATASETS AND SCENARIOS

Datasets. Three transmission line inspection datasets are adopted in our experiments.

- **Inspection-A** is a transmission line inspection dataset collected from city A. It contains 1680 training images and 336 validation images. The images are captured through the camera carried by UAV in normal weather conditions. The collected images retain the size of the original images, and each picture contains multiple detection objects. There are seven categories with instance labels in Inspection-A, which are glass insulator, polymer insulator, shockproof, clamps, birds nests, and fault insulator.

- **Rain Inspection-A** is built upon the images in the Inspection-A, because labeling rainy days data with accurate bounding box annotations costs prohibitively. Therefore, we followed [70] to synthesize Rain scenes by using Photoshop,\(^1\) so that the bounding box annotations of Inspection-A can be reused. The bounding box annotations of objects and division of the dataset are inherited from the Inspection-A dataset. We use this dataset for weather adaptability experiments.

- **Inspection-B** is a transmission line inspection dataset collected from city B. It contains 3168 training images and 792 validation images. It differs from Inspection-A in that the images come from different real-world and that the UAV video camera model for which the images were collected is different [71]. In addition, these images are clipped, and most of them contain only one object, so the size of the objects in different images is similar. There are two categories with instance labels in Inspection-B, which are glass insulator, polymer insulator.

Figure 4 shows some samples of these datasets.

Scenarios. We evaluate the proposed method on two domain adaptation scenarios:

- **Cross-Weather** (Inspection-A \( \rightarrow \) Rain Inspection-A). It aims to adapt object detection to different weather conditions because the scenes are visually different as weather situations transform. Whether a detection model can work reliably in various weather situations is essential for intelligent transmission line inspection. So we study the performance of detecting objects when we transfer the scene from normal to rainy. During the training phase, we use the training set of Inspection-A and Rain Inspection-A as source and target domain respectively. The results of detection are reported in the validation set of Rain-Inspection-A.

- **Cross-Camera** (Inspection-A \( \rightarrow \) Inspection-B and Inspection-B \( \rightarrow \) Inspection-A). The images in different datasets are captured by different cameras. However, different cameras’ settings result in different image quality/resolution. Therefore, even with similar weather conditions, the images have domain shift among them [71]. Besides, the varied gap in scale, size and category distribution among different datasets will affect the domain shift. In this paper, we focus on researching adaptation between two real-world datasets. To simulate this adaptation, we use the training set of Inspection-A as the source domain and the training set of Inspection-B as the target domain. Besides, we also did an Inspection-B \( \rightarrow \) Inspection-A. It should be noted that instance labels of Inspection-B only includes two classes. As a result, in the cross-camera scene task, the adaptive object detection category is the glass insulator and polymer insulator.

B. IMPLEMENTATION DETAILS

Our detection model employed Faster R-CNN [29] and ROIAlign [30] as base detection architecture in all experiments. The backbone of our model is initialized using the pre-trained weights of VGG-16 [66] on ImageNet [72]. The shorter sides of all input images are resized to 600 pixels. At each iteration, the batch has two images, one labeled sample belonging to the source domain, and the other unlabeled sample belonging to the target domain. We evaluate the adaptation detection performance by reporting mean average precision(mAP) with an IoU threshold of 0.5. We fine-tune the detection model by a learning rate of

\(^1\)http://www.photoshopessentials.com/photo-effects/rain/
$1 \times 10^{-3}$ for 50k iterations and then lessen the learning rate to $1 \times 10^{-4}$ for the last 20k iterations. We adopt mini-batch stochastic gradient descent (SGD) with the momentum of 0.9 and the weight decay of $5 \times 10^{-4}$ to all parameters updated. For Inspection-A $\rightarrow$ Rain-Inspection-A, we set $\lambda = 1$ in Eqn. 6. For Inspection-A $\rightarrow$ Inspection-B and Inspection-B $\rightarrow$ Inspection-A, we set $\lambda = 0.01$ in Eqn. 6. We adopt PyTorch [73] to implement our experiments.

We compare our model with several state-of-the-art domain adaptive object detection algorithms: (1) Source only stands for the Faster-RCNN [29] model, which is trained on the source data. Note that there is no domain adaptation in this model. (2) Domain adaptive Faster-RCNN (DAF) [17], which reduces the divergence of cross-domain distribution at both image-level and instance-level. The image-level domain discriminator is trained on the feature map after the base convolutional layers (block5) for the image-level alignment, while the instance-level domain classifier is trained on instance-level RoI features for aligning the instance-level features, and the consistency loss regularizes between image-level and instance-level domain discriminators to learn a robust model. (3) Categorical Regularization for Domain Adaptive Object Detection [18] proposed a categorical regularization framework to focus on aligning object-related local regions and hard aligned instances by utilizing image-level categorical regularization (ICR) and the categorical consistency regularization (CCR) between image-level and instance-level predictions. (4) Multi Adversarial Faster-RCNN (MAF) [14], which adopts multi-level feature alignment method on the image-level features and detection prediction as a constraint condition for instance-level alignment. (5) iFAN: Image-Instance Full Alignment [16], which conducts image-level features alignment in multiple intermediate layers. For instance-level alignment, it uses predicted bounding boxes to extract feature of instance, with metric learning to learn cross-domain category correlations. In addition, it also adopts conditional adversarial domain adaptation.

### C. RESULTS

In order to clearly compare our results on image-level alignment, we also provide the experimental results of DAF [17], DAF-ICR [18], MAF [14], and iFAN [16] with only image-level adaptation. To research the influence of the proposed method on each module, we design ablation experiments on the Inspection-A $\rightarrow$ Rain-Inspection-A task. Our baseline method is referred to as 3DA, which is blocks 3, 4, and 5 of the Faster R-CNN model integrated with three domain adversarial subnetworks.

**Cross-Weather.** Table 1 shows the performance of adaptation from Inspection-A $\rightarrow$ Rain Inspection-A. We achieve an mAP of 79.7%, which is 30.3% higher than the source only model. It is worth noting that our method only completes cross-domain alignment in the image-level. Among all results in Table 1, the result of the proposed method is the best. Our method outperforms iFAN [16] and MAF [14] by 1.9% and 1.6%, respectively. The compelling results demonstrate that, with the image-level adaptation of multi-scale context enhancement alignment and multi-scale consistency regularization, our model can learn more discriminative representation in the target domain.

**Cross-Camera.** As shown in table 2, we illustrate the performance comparison in both adaptation detection tasks. The proposed method achieves an mAP of 79.7%, which is 30.3% higher than the source only model. It is worth noting that our method only completes cross-domain alignment in the image-level. Among all results in Table 1, the result of the proposed method is the best. Our method outperforms iFAN [16] and MAF [14] by 1.9% and 1.6%, respectively. The compelling results demonstrate that, with the image-level adaptation of multi-scale context enhancement alignment and multi-scale consistency regularization, our model can learn more discriminative representation in the target domain.

### D. ANALYSIS

**Ablation Study.** To research the influence of the proposed method on adaptation object detection, we design ablation experiments on the Inspection-A $\rightarrow$ Rain Inspection-A task.

| Method               | Glass Insulator | Polymer Insulator | Shockproof | Clamps | Birds Nests | Fault Insulator | mAP  |
|----------------------|-----------------|-------------------|------------|--------|------------|-----------------|------|
| Source Only [29]     | 74.2%           | 64.8%             | 25.8%      | 20.1%  | 47.6%      | 63.9%           | 49.4%|
| DAF(CVPR’18) [17]    | 85.6%           | 84.6%             | 41.5%      | 36.9%  | 55.0%      | 77.6%           | 63.5%|
| DAF (image-level)    | 86.0%           | 82.6%             | 37.3%      | 40.7%  | 52.4%      | 81.3%           | 63.4%|
| DAF-ICR-CCR (CVPR’20)| 85.5%           | 85.06%            | 41.0%      | 40.1%  | 60.2%      | 80.7%           | 65.4%|
| DAF-ICR (image-level)| 84.1%           | 83.9%             | 39.6%      | 39.9%  | 58.6%      | 78.3%           | 64.1%|
| MAF (ICCV’19) [14]   | 89.0%           | 88.5%             | 46.0%      | 58.9%  | 95.2%      | 90.8%           | 78.1%|
| MAF (image-level)    | 90.3%           | 86.7%             | 43.1%      | 52.7%  | 93.3%      | 90.1%           | 76.0%|
| iFAN (AAAI’20) [16]  | 89.5%           | 90.0%             | 43.3%      | 57.5%  | 96.2%      | 90.6%           | 77.8%|
| iFAN (image-level)   | 89.9%           | 86.4%             | 42.4%      | 51.5%  | 94.8%      | 89.6%           | 75.8%|
| 3DA                  | 89.1%           | 79.3%             | 44.6%      | 43.8%  | 80.9%      | 90.9%           | 71.4%|
| Ours                 | 90.6%           | 90.5%             | 46.5%      | 60.6%  | 99.6%      | 90.5%           | 79.7%|
TABLE 2. Cross-Camera Adaptation: Results on Inspection-B, using models trained on Inspection-A.

| Method                  | Glass Insulator | Polymer Insulator | mAP  |
|-------------------------|-----------------|-------------------|------|
| Source                  | 38.4%           | 44.7%             | 41.5%|
| DAF (CVPR'18)           | 38.0%           | 47.8%             | 42.9%|
| DAF (Image-level)       | 39.9%           | 46.3%             | 43.1%|
| DAF-ICR-CCR(CVPR'20)    | 39.5%           | 46.7%             | 43.1%|
| DAF-ICR (image-level)   | 39.2%           | 46.5%             | 42.9%|
| MAF (ICCV'19)           | 39.8%           | 48.3%             | 44.0%|
| MAF (Image-level)       | 39.2%           | 48.2%             | 43.7%|
| iFAN (AAAI'20)          | 40.2%           | 48.3%             | 44.3%|
| iFAN (Image-level)      | 38.9%           | 47.0%             | 43.4%|
| 3DA                     | 38.8%           | 46.7%             | 42.7%|
| Ours                    | 40.2%           | 50.3%             | 45.3%|

TABLE 3. Cross-Camera Adaptation: Results on Inspection-A, using models trained on Inspection-B.

| Method                  | Glass Insulator | Polymer Insulator | mAP  |
|-------------------------|-----------------|-------------------|------|
| Source                  | 38.8%           | 47.2%             | 43.0%|
| DAF (CVPR'18)           | 49.3%           | 49.9%             | 49.6%|
| DAF (Image-level)       | 49.2%           | 47.9%             | 48.5%|
| DAF-ICR-CCR (CVPR'20)   | 50.6%           | 49.7%             | 50.2%|
| DAF-ICR (image-level)   | 50.1%           | 49.5%             | 49.8%|
| MAF (ICCV'19)           | 54.9%           | 48.8%             | 51.8%|
| MAF (Image-level)       | 52.3%           | 48.9%             | 50.6%|
| iFAN (AAAI'20)          | 53.9%           | 50.2%             | 52.1%|
| iFAN (Image-level)      | 51.1%           | 49.2%             | 50.1%|
| 3DA                     | 51.5%           | 46.8%             | 50.0%|
| Ours                    | 55.2%           | 50.7%             | 52.9%|

The study evaluates several variants of the proposed method: the baseline is 3DA, which denotes that Multi-Scale Fusion Feature Alignment and Multi-Scale Consistency Regularization are not adopted in the adaptation detection process; Ours(w/o MSFA), which means that we remove the Multi-Scale Fusion Feature Alignment from the network but preserve the consistency regularization; Ours(w/o MSCR), which means there is no Multi-Scale Consistency Regularization in the training process. The results are shown in table 4. The mAP of the proposed model (w/o MSFA) is significantly improved compared with 3DA, which is because the domain invariant feature learning are consistent among different layers. Table 4 prove that the MSFA module that uses multi-scale fusion function for hierarchical domain adaptation is very useful. In a nutshell, this paper’s motivation is correct, and each of these methods significantly improves performance.

Error analysis of highest confident detections. To further validate the effect of the proposed method for adaptation object detection, we analyze the errors in the most confident detections. The experiment is conducted with models of Source Only, MAD, iFan, and Ours in the Cross-Weather scene. We follow [17], [74] to categorize the detections into three analysis types: Correct (The overlap between the prediction and the ground-truth ≥ 0.5), Mislocalization (0.3 ≤ The overlap between the prediction and the ground-truth ≤ 0.5), and Background (The overlap between the prediction and the ground-truth ≤ 0.3, which means the prediction is a false positive). We choose the top-N detections score for each category, where N is the number of instances in this category. Figure 5 shows the results of each analysis type for all categories. Compared to Source Only, MAD, iFan, the proposed method significantly increases the percentage of correct prediction (green color) and decreases the false positives (red and gray colors).

Parameter Sensitivity. We check the sensitivity of our method hyper-parameter λ which is used to balance the detection loss and domain adaptation loss. Figure 6(a) demonstrates the parameter sensitivity of our method on all tasks Inspection-A → Rain-Inspection-A, Inspection-A → Inspection-B and Inspection-B → Inspection-A. We plot detection accuracy w.r.t different values of λ in Figure 6(a), which indicates that λ ∈ [0.01, 1.0] can be optimal parameter values where our method generally does much better than the baselines.

TABLE 4. Ablation of the proposed method on Inspection-A → Rain Inspection-A.

| Method               | Glass Insulator | Polymer Insulator | Shockproof | Clamps | Birds Nests | Fault Insulator | mAP  |
|----------------------|-----------------|-------------------|------------|--------|------------|-----------------|------|
| 3DA                  | 89.1%           | 79.3%             | 44.6%      | 43.8%  | 80.9%      | 90.9%           | 71.4%|
| Ours w/o MSFA        | 89.8%           | 87.0%             | 48.0%      | 57.5%  | 96.4%      | 87.9%           | 77.8%|
| Ours w/o MSCR        | 89.8%           | 86.2%             | 44.5%      | 56.1%  | 99.3%      | 89.8%           | 77.6%|
| Ours                 | 90.6%           | 90.5%             | 46.5%      | 60.6%  | 99.6%      | 90.5%           | 79.7%|
Convergence and Time Complexity. We also empirically check the convergence property of our method. We used Maximum Mean Discrepancy (MMD) [42] to measure the distribution discrepancy of learned features during the training phase. Figure 6(b) shows that distribution distance decreases and converges within several iterations.

Example of Detection Results. Figure 7 shows some visualization results of detection on Cross-Weather scenarios. Even when the image style and the instance scale of source and target domain are widely different, our proposed method produces more accurate bounding box predictions under complex scenarios and domain shifts.

V. CONCLUSION
In order to reduce the domain shift caused by multi-scale in the cross-scene power transmission line inspection task. In this paper, we proposed a multi-scale feature enhanced domain adaptive object detection for cross-scene transmission lines inspection. To alleviate the influence domain shift caused by scales of instances, we develop a novel approach for detector adaptation based on an adversarial framework, which includes multi-scale fusion feature alignment and multi-scale consistency regularization. In the MSFA module, we utilizing context information conveyed from other levels to construct a multi-scale fusion feature and learning more similar representation characteristics of different scales objects by aligning fusion features. Additionally, we propose the multi-scale consistency regularization.
to make the domain invariant feature learning consistent among different layers. To verify the efficacy of the proposed method and prepare for further research, we construct three transmission lines inspection datasets of different scenes. We evaluate the proposed method on cross-weather and cross-camera adaptation scenarios. The experimental results testify the effectiveness of our model on object detection of cross-scene transmission lines.

REFERENCES

[1] B. Mann and I. Morrison, “Digital calculation of impedance for transmission line protection,” IEEE Trans. Power App. Syst., vol. PAS-90, no. 1, pp. 270–279, Jan. 1971.

[2] H. Wang and W. W. L. Keerthipala, “Fuzzy-neuro approach to fault classification for transmission line protection,” IEEE Trans. Power Del., vol. 13, no. 4, pp. 1093–1104, Oct. 1998.

[3] S. Danbhare, S. A. Soman, and M. C. Chandorkar, “Adaptive current differential protection schemes for transmission-line protection,” IEEE Trans. Power Del., vol. 24, no. 4, pp. 1832–1841, Oct. 2009.

[4] F. Tian, Y. Wang, and L. Zhu, “Power line recognition and tracking method for UAVs inspection,” in Proc. IEEE Int. Conf. Inf. Autom., Aug. 2015, pp. 2136–2141.

[5] G. Zhou, J. Yuan, I.-L. Yen, and F. Bastani, “Robust real-time UAV based power line detection and tracking,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2016, pp. 744–748.

[6] H. Jiang, X. Qiu, J. Chen, X. Liu, X. Miao, and S. Zhuang, “Insulator fault detection in aerial images based on ensemble learning with multi-level perception,” IEEE Access, vol. 7, pp. 61797–61810, 2019.

[7] X. Liu, J. Jiang, J. Chen, S. Zhuang, and X. Miao, “Insulator detection in aerial images based on faster regions with convolutional neural network,” in Proc. IEEE 14th Int. Conf. Control Autom. (ICCA), Jun. 2018, pp. 1082–1086.

[8] G. F. Gao, J. Wang, Z. Kong, J. Wu, N. Feng, S. Wang, P. Hu, Z. Li, H. Huang, and J. Li, “Recognition of insulator explosion based on deep learning,” in Proc. 14th Int. Conf. Comput. Wavelet Act. Media Technol. Inf. Process. (ICCWAMITIP), Dec. 2017, pp. 79–82.

[9] X. Miao, X. Liu, J. Chen, S. Zhuang, J. Fan, and H. Jiang, “Insulator detection in aerial images for transmission line inspection using single shot multibox detector,” IEEE Access, vol. 7, pp. 9945–9956, 2019.

[10] Z. A. Siddiqui, R. Park, S.-W. Lee, N.-J. Jung, M. Choi, C. Lim, and J.-H. Seo, “Robust powerline equipment inspection system based on a convolutional neural network,” Sensors, vol. 18, no. 11, p. 3837, Nov. 2018.

[11] R. Gopalani, R. Li, and R. Chellappa, “Domain adaptation for object recognition: An unsupervised approach,” in Proc. Int. Conf. Pattern Recognit., Nov. 2011, pp. 999–1006.

[12] S. Jialin Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[13] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang, “Multi-level domain adaptive learning for cross-domain detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019.

[14] Z. He and L. Zhang, “Multi-adversarial faster-RCNN for unrestricted object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6668–6677.

[15] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Strong-weak distribution alignment for adaptive object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6956–6965.

[16] C. Zhang, X. Han, W. Huang, and M. R. Scott, “IFAN: Image-instance full networks for adaptive object detection,” 2020, arXiv:2003.04132. [Online]. Available: http://arxiv.org/abs/2003.04132

[17] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3339–3348.

[18] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring categorical regularization for domain adaptive object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11724–11733.

[19] P. Zhou, B. Ni, J. Hu, and Y. Xu, “Scale-transferable object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 528–537.

[20] L. Wei, W. Cui, Z. Hu, H. Sun, and S. Hou, “A single-shot multi-level feature reused neural network for object detection,” Vis. Comput., vol. 3, pp. 1–10, Jan. 2020.
BING LUO: received the M.S. degree in electrical engineering from Chongqing University, China, in 1990, and the Ph.D. degree in electrical engineering from Tsinghua University, China, in 1996. He is currently a Professorate Senior Engineer with the Electrical Power Research Institute of China South Grid, China. His research interests include high-voltage insulation test technology and insulation detection of electrical power equipment.

TINGTING WANG: received the B.S. and M.S. degrees in electrical engineering from the Harbin University of Science and Technology, China, in 2013 and 2016, respectively. She is currently an Engineer with the Electrical Power Research Institute of China South Grid, China. Her research interests include high-voltage test technology and insulator detection.