Unseen-Material Few-Shot Defect Segmentation with Optimal Bilateral Feature Transport Network

Dexing Shan, *Student Member, IEEE*, Yunzhou Zhang, *Member, IEEE*, Sonya A. Coleman, Dermot Kerr, Shitong Liu, and Ziqiang Hu

***Abstract*— Industrial defect segmentation is important to ensure product quality and production safety. The main challenges in industrial applications are insufficient defect samples, large intra-class variation, and the interference of background information. However, most current texture defect segmentation methods rely on large-scale datasets and can only deal with one specific type of texture defect, which reduces the application efficiency and application scope of defect segmentation algorithms. To this end, we propose an optimal bilateral feature transport network (OBFTNet) for few-shot texture defect segmentation, which can accurately segment texture defects in multiple unseen materials (domains), such as steel, wood, leather, etc. OBFTNet can perform bilateral prediction for background and defect regions of unseen material by dynamically predicting task-specific semantic correspondences conditioned on a small guidance set. Specifically, we introduce background images (defect-free images) as supplementary learning information for reverse prediction and model the semantic correspondence between the guidance (support and background images) and the query images in few-shot segmentation as an optimal bilateral feature transport problem and generate a set of optimal bilateral correlation tensors. Using 4D and 2D convolutions the model gradually reduces optimal bilateral correlation tensors to precise segmentation masks. Experimental results show that our proposed method outperforms several state-of-the-art techniques with very few labeled samples and the method generalizes well to industrial defects on unseen materials.**

***Index Terms*— Few-shot segmentation, optimal transport, texture defect segmentation, cross-domain**

I. Introduction

IN the manufacturing industry, the various stages of a material's complex manufacturing processes can cause a range of defects on the product surface. Defect segmentation can accurately determine the pixel-level position of defects, which plays a significant role in ensuring product quality in the manufacturing process. Visual defect segmentation is a non-contact, sensor-based technology that has the advantages of flexibility and accuracy and in recent years deep learning has significantly improved the performance of this task. Inspired by classic segmentation networks [1], [2], the main ideas to improve defect segmentation are to fuse multi-scale features [3]–[5] or introduce an attention mechanism [3], [6] such that the model can accurately segment defects of different sizes. These methods have achieved good performance for defect segmentation. However, such methods often suffer from performance degradation for unseen classes as they rely on large-scale labeled datasets.

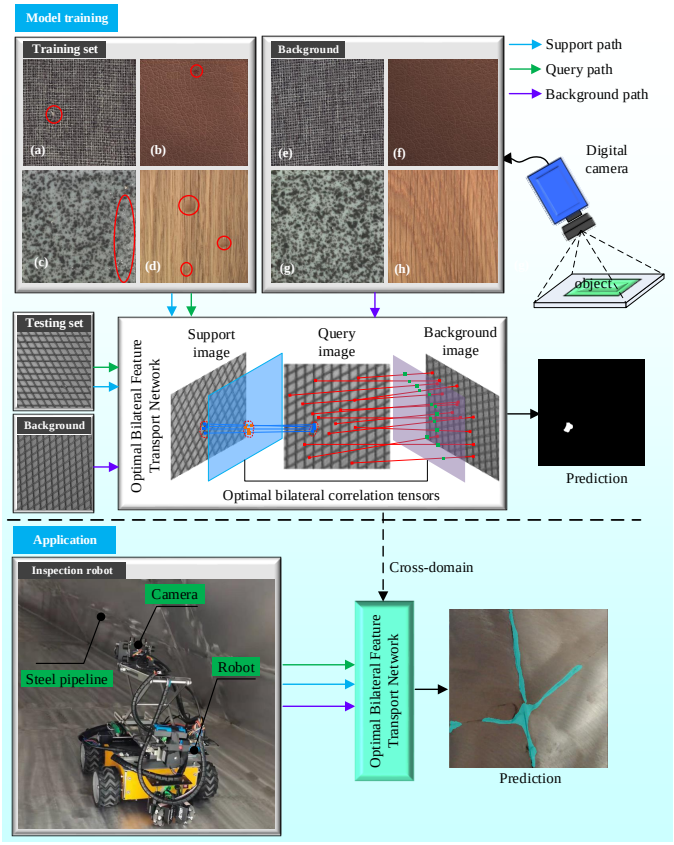

Fig. 1. Overview of the proposed visual inspection technology for few-shot unseen material texture defect segmentation. (a) Defects on the carpet. (b) Defects on the leather. (c) Defects on the tile. (d) Defects on the wood. Images (e)-(h) are the background images corresponding to the training set images.

Manuscript received June 1, 2022; revised August 9, 2022; accepted September 27, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61973066, in part by the Major Science and Technology Projects of Liaoning Province under Grant 2021JH1/10400049, in part by the Foundation of Key Laboratory of Aerospace System Simulation under Grant 6142002200301, and in part by the Foundation of Key Laboratory of Equipment Reliability under Grant WD2C20205500306. (Corresponding author: Yunzhou Zhang.)

Dexing Shan, Yunzhou Zhang, Shitong Liu, and Ziqiang Hu are with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: 2010731@stu.neu.edu.cn; zhangyunzhou@mail.neu.edu.cn; 2100789@stu.neu.edu.cn; 2100751@stu.neu.edu.cn

Sonya A. Coleman and Dermot Kerr are with the Intelligent Systems Research Centre, Ulster University, Londonderry BT48 7JL, U.K. (e-mail: sa.coleman@ulster.ac.uk; d.kerr@ulster.ac.uk).

Although some methods [7]–[9] can segment defects from images without any labels, a large-scale defect dataset is still required [10]. In particular, the dependence of models on large-scale datasets in industrial applications also reduces the deployment efficiency of defect detection algorithms. Addressing the problem of insufficient samples of manufacturing texture defects, we propose a few-shot segmentation method with an optimal bilateral feature transport network (OBFTNet). We can use OBFTNet to segment unseen material defects thereby solving the issue of excessive dependence of defect segmentation algorithms on large-scale defect datasets. We apply this defect segmentation method to a steel surface inspection robot, as shown in Fig. 1. In the manufacturing industry, the segmentation of unseen texture defects has four main challenges: insufficient defect samples, scale changes, complex background information interference, and texture shift. Some common manufacturing defects are shown in Fig. 1(a)-(d).

To deal with the challenge of insufficient defect samples, we apply few-shot learning for texture defect segmentation. Few-shot segmentation was introduced by Shaban *et al.* [11], which is used to segment objects in a query image when only a few labeled support samples are available. At present, few-shot semantic segmentation has been applied in the fields of video object segmentation [12], medical image segmentation [13], and industrial defect segmentation [14]. There are two widely used methods for few-shot semantic segmentation: prototype learning and affinity learning. Prototype learning techniques compress the foreground object in the support image into one or several prototype feature vectors [15]–[20], and then find the pixel positions of similar features in a query image to segment the desired target. Typically as only one prototype is obtained from the support feature, this model inevitably loses context information [19]. The affinity learning networks [21], [22] establish dense pixel-level correlations between the support and query images, which preserve spatial information better than prototype learning approaches. To date there are few current studies on few-shot texture defect segmentation. In this work, we apply the theoretical research of few-shot learning in practice to complete the task of few-shot texture defect segmentation. Aiming to solve the problem of limited appearance information of texture defects, we adopt affinity learning for few-shot texture defect segmentation to fully mine and utilize limited defect information, as shown in Fig. 1(Optimal bilateral feature transport network).

To deal with the challenge of scale changes, our model utilizes a rich set of features in the middle layers of the backbone network to capture multi-scale semantic information between support and query images, which can improve the model's ability to perceive defects from different scales.

To deal with the challenge of complex background interference, we introduce the often-ignored background image (see Fig. 1(e)-(h)) to establish a bilateral (foreground and background) semantic guidance mechanism to improve the discriminative capability of the model. For clarity, we refer to the background image and the support image together as the **guidance image**.

To deal with the challenge of the domain gap and texture shift between different materials, we establish a precise semantic correspondence between the query and the guidance images by adopting optimal transport (OT) [23]. This alleviates the many-to-one feature matching problem, to activate full object extent in a query image. When cross-domain to other material, this approach can adaptively generate task-specific semantic correspondences conditioned on a smaller guidance image, overcoming domain gaps and texture shift issues.

In our proposed few-shot texture defect segmentation algorithm, there are four key contributions:

1) To the best of our knowledge, it is the first time that the few-shot learning theory is used to segment the defects in unseen domains.
2) We introduce background images as supplementary learning information and improve the model's ability to distinguish complex backgrounds via reverse prediction so that the classifier can focus more on learning foreground objects.
3) We model the semantic correspondence between the guidance and the query images, in few-shot segmentation, as an optimal bilateral transport problem and adaptively generate task-specific semantic correspondence to guarantee the model generalization to unseen materials.
4) Extensive experiments using multiple popular backbone networks show that our proposed method exceeds the performance of state-of-the-art techniques. Furthermore, it performs well under a more challenging yet practical cross-domain setting.

II. RELATED WORK

A. Defect Segmentation

Defect semantic segmentation methods based on supervised learning rely on large-scale labeled datasets for training. PGA-Net [3] proposes pyramid features fusion and global context attention to achieve pixel-level defect detection. Cao *et al.* [4] overcomes the problem of the small number of samples in deep learning by using a two-stage network structure combining a segmentation network and a decision network to realize defect semantic segmentation. Song *et al.* [5] proposed a pixel-level segmentation network based on multi-scale feature fusion for defect semantic segmentation. The defect semantic segmentation methods [9], [24], [25] based on unsupervised learning further overcome the dependence of defect semantic segmentation networks on pixel-level labels. The above technologies have achieved good performance in industrial defect inspection, but they still have the problem of relying on a large number of samples and poor generalization to unseen domains, which increases the development cost and development cycle of defect detection algorithms.

B. Optimal Transport

Optimal transport provides a way to infer the correspondence between two distributions. SCOT [26] uses semantic correspondence as an optimal transport problem to establish a global optimal matching probability between two images. SuperGlue [27] uses the Sinkhorn algorithm for the score matrix

to obtain the optimal allocation matrix and then obtains the best matching key points according to the allocation matrix. UNITE [28] introduces optimal transport for accurate feature alignment and faithful style control in image translation. To the best of our knowledge, we are the first to introduce an optimal transport framework to few-shot semantic segmentation.

C. Few-shot Segmentation

At present, there are two widely used methods for few-shot semantic segmentation: prototype learning and affinity learning. PFENet [16] proposed the prior guided feature enrichment network, which uses a prior mask generation method without training to ensure generalization. Work such as PMMs [18] and ASGNet [19] aimed to generate multiple prototypes. Based on the traditional training process support-query pairs, MiningFSS [29] introduces an additional mining branch to mine latent new classes through transferable sub-clusters. These prototype-based methods need to compress the support image information into different prototypes. One major drawback of this approach is that it leads to the loss of spatial information. Alternatively, PGNet [21], HFA [22], GuidedNet [30], co-FCN [31], VAT [32], and HSNet [33] established pixel-level correlation between support and query images based on affinity learning. Although affinity learning can maintain more spatial information, it is susceptible to noise information [19]. Therefore, we use the optimal transport theory to optimize affinity learning to alleviate the impact of noisy and incorrect matching problems on affinity learning.

III. METHOD

In this section, we present a novel few-shot defect segmentation architecture, Optimal Bilateral Feature Transport Networks (OBFTNet), which consists of a multi-scale feature extraction network, an optimal bilateral feature transport module, and a correlation analysis module, as shown in Fig. 2. Further detail is provided in the following subsections.

A. Task Setting

The aim of the few-shot texture defect segmentation task is to train a model to segment defect regions in a query image by referencing only a few annotated support images that contain the same defect categories. It is notable that, during the inference phase, defect categories and materials on the testing set are completely unseen during training. Specifically, given two sets: a training set $D_{\text{train}} = (X_{\text{S}}, X_{\text{Q}}, X_{\text{B}}, Y_{\text{S}}, Y_{\text{Q}})$ and a testing set $D_{\text{test}} = (X_{\text{S}}, X_{\text{Q}}, X_{\text{B}}, Y_{\text{S}}, Y_{\text{Q}})$, where the D_{train} is used for training the model and the D_{test} is for evaluation. The categories of the two sets do not intersect $(D_{\text{train}} \cap D_{\text{test}} = \emptyset)$. X indicates the texture image and Y denotes the ground-truth binary mask. Subscripts S, Q, and B represent the support image, the query image, and the background image, respectively. Following the conventional few-shot segmentation [15]–[19], [33]–[35], our proposed OBFTNet requires that the support image, query image, and background image come from the same material, and the support image and query image contain the same defect class. Given a k-shot learning task, the model randomly selects a defect class among the defect classes to be tested. The system randomly chooses a query image X_{Q}, a background image X_{B}, K support images X_{S}^k, and corresponding support masks Y_{S}^k from the predefined database according to the chosen defect class, where $k \in \{1, \ldots, K\}$.

B. Multi-scale Feature Extraction

Inspired by recent semantic matching methods [36], our model exploits a rich set of multi-scale features from intermediate layers of the backbone network to capture the multi-scale semantic similarity between guidance and query images. Given a set of support, query, and background images, X_{S}, X_{Q}, $X_{\text{B}} \in \mathbb{R}^{3\times H\times W}$, a sequence of feature maps $\left\{(\mathbf{F}_l^{\text{s}}, \mathbf{F}_l^{\text{q}}, \mathbf{F}_l^{\text{b}})\right\}_{l\in\mathcal{L}_p}$ can be generated through a weight sharing backbone network, where $\mathcal{L}_p$ is a set of CNN layer indices $\{1, \ldots, L\}$ at some pyramidal layer p, $p \in \{1, \ldots, P\}$, $\mathbf{F}_l^{\text{s}}$, $\mathbf{F}_l^{\text{q}}$, and $\mathbf{F}_l^{\text{b}}$ denote support, query, and background feature maps at l-th level of pyramidal layer p, respectively.

We use the support mask $Y_{\text{S}} \in \{0,1\}^{H_p\times W_p}$ to mask out the background information in the support feature maps $\mathbf{F}_l^{\text{s}} \in \mathbb{R}^{C_l\times H_p\times W_p}$ as

$$\hat{\mathbf{F}}_l^{\text{s}} = \mathbf{F}_l^{\text{s}} \otimes {Y_{\text{S}}}^p \tag{1}$$

where $\otimes$ is the broadcasting element-wise multiplication, ${Y_{\text{S}}}^p \in \mathbb{R}^{H_p\times W_p}$ is the support mask at pyramidal layer p, C_l is the number of channels at the CNN layer l, H_p and W_p are the height and width of the feature maps at pyramidal layer p, respectively.

C. Optimal Bilateral Feature Transport Module

As shown in Fig. 3, we first calculate foreground correlation maps and background correlation maps by cosine similarity to establish initial bilateral semantic correspondence from the guidance image to the query image, then model it as an optimal transport problem. Finally, we obtain the pyramid optimal bilateral correlation tensors.

1) Correlation map: At pyramidal layer p, we obtain the initial foreground correlation map $\mathbf{S}_l^{\text{s}}$ by calculating the cosine similarity between query features $\mathbf{F}_l^{\text{q}}$ and foreground support features $\hat{\mathbf{F}}_l^{\text{s}}$ as

$$\mathbf{S}_l^{\text{s}} = \frac{\mathbf{F}_l^{\text{q}} \cdot \hat{\mathbf{F}}_l^{\text{s}}}{\|\mathbf{F}_l^{\text{q}}\| \left\|\hat{\mathbf{F}}_l^{\text{s}}\right\|} \in \mathbb{R}^{M_p\times N_p} \tag{2}$$

where $\|\cdot\|$ is the L2 normalization, $M_p = N_p = H_p \times W_p$. Similarly, the initial background correlation map $\mathbf{S}_l^{\text{b}}$ is obtained by calculating the cosine similarity between query features $\mathbf{F}_l^{\text{q}}$ and background features $\mathbf{F}_l^{\text{b}}$ as

$$\mathbf{S}_l^{\text{b}} = \frac{\mathbf{F}_l^{\text{q}} \cdot \mathbf{F}_l^{\text{b}}}{\|\mathbf{F}_l^{\text{q}}\| \left\|\mathbf{F}_l^{\text{b}}\right\|} \in \mathbb{R}^{M_p\times N_p} \tag{3}$$

We collect $\mathbf{S}_l^{\text{s}}$ and $\mathbf{S}_l^{\text{b}}$ with the same spatial size and denote the subset as $\{\mathbf{S}_l^{\text{s}}\}_{l=1}^{L}$ and $\{\mathbf{S}_l^{\text{b}}\}_{l=1}^{L}$, respectively. Finally, all correlation maps in $\{\mathbf{S}_l^{\text{s}}\}_{l=1}^{L}$ and $\{\mathbf{S}_l^{\text{b}}\}_{l=1}^{L}$ are concatenated along channel domain to form $\mathbf{S}_p^{\text{s}} \in \mathbb{R}^{L\times M_p\times N_p}$ and $\mathbf{S}_p^{\text{b}} \in$

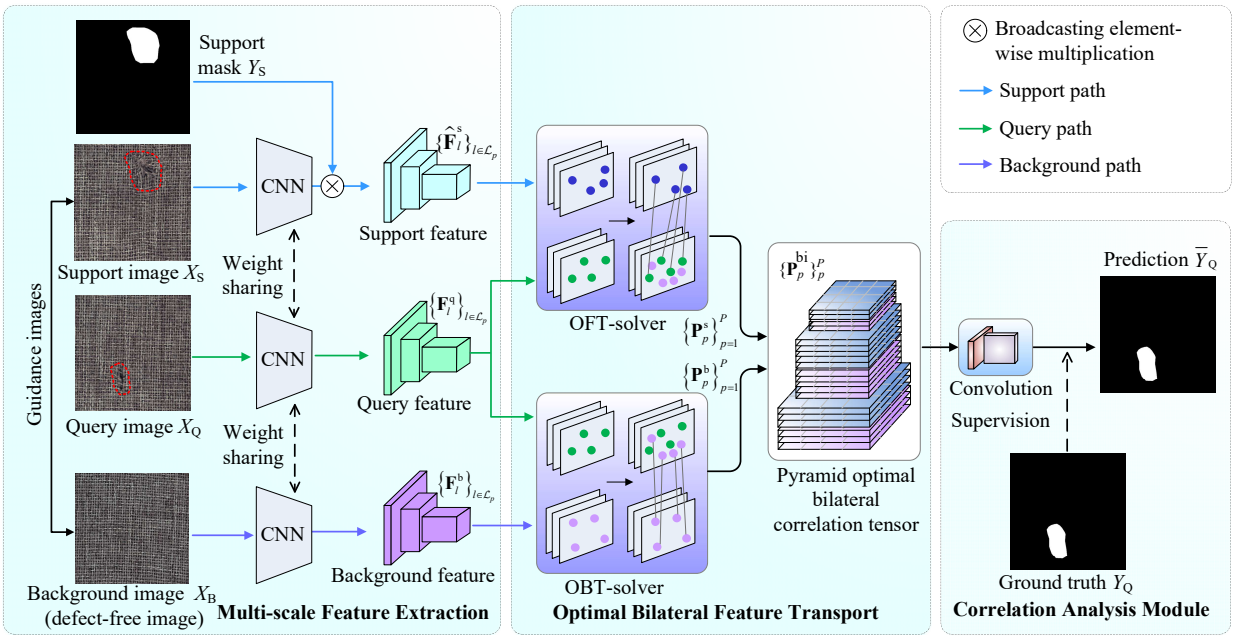

Fig. 2. Overall architecture of the proposed network which consists of three main parts: Multi-scale Feature Extraction, Optimal Bilateral Feature Transport, Correlation Analysis Module.

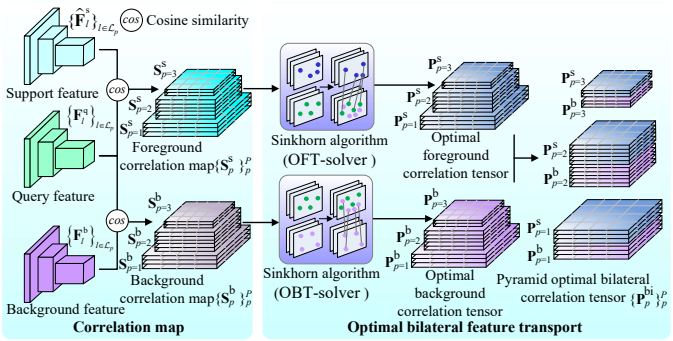

Fig. 3. Optimal Bilateral Feature Transport Module.

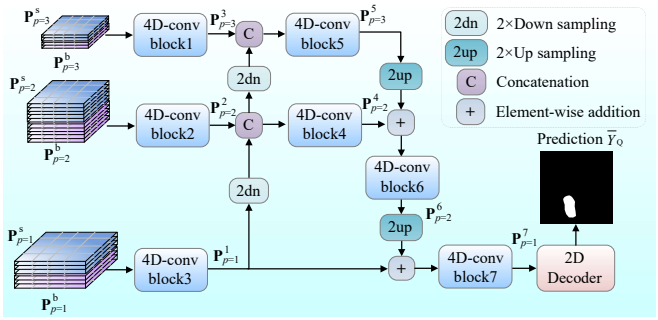

Fig. 4. Correlation Analysis Module.

$\mathbb{R}^{L\times M_p\times N_p}$, where L is the number of CNN layers at pyramid layer p.

Given P pyramidal layers, we collect the correlation maps $\mathbf{S}_p^{\mathrm{s}}$ and $\mathbf{S}_p^{\mathrm{b}}$ of each pyramidal layer p to obtain a set of multi-scale foreground correlation maps $\left\{\mathbf{S}_p^{\mathrm{s}}\right\}_{p=1}^{P}$ and a set of multi-scale background correlation maps $\left\{\mathbf{S}_p^{\mathrm{b}}\right\}_{p=1}^{P}$.

In this process, correlation maps are obtained by calculating the cosine similarity of the two sets of local features, but the pairwise matching scores of each location are calculated individually without considering any mutual relationship. However, due to the large intra-class variation and the interference of background, most of the information in the correlation map is matching noise. Therefore, we use optimal transport theory to optimize the initial bilateral correlation maps ($\left\{\mathbf{S}_p^{\mathrm{s}}\right\}_{p=1}^{P}$ and $\left\{\mathbf{S}_p^{\mathrm{b}}\right\}_{p=1}^{P}$) to alleviate the impact of noisy information and incorrect matching problems on affinity learning.

2) Solving OT with the Sinkhorn algorithm: Our goal is to establish an accurate semantic correspondence between guidance and query images. The optimal bilateral feature transport (OFT-solver and OBT-solver) is the core of OBFTNet, which adaptively generates task-specific semantic correspondences conditioned on a small guidance set, overcoming the domain gap and texture-shift issues. In this way, limited guidance information can be used reasonably, and more pixels on the foreground object of the query image can be activated.

The solution to the above optimization problem corresponds to the optimal transport between $\mathbf{F}_l^{\mathrm{q}}$ and $\hat{\mathbf{F}}_l^{\mathrm{s}}$ with a foreground correlation map $\mathbf{S}_l^{\mathrm{s}}$, and the optimal transport between $\mathbf{F}_l^{\mathrm{q}}$ and $\mathbf{F}_l^{\mathrm{b}}$ with a background correlation map $\mathbf{S}_l^{\mathrm{b}}$. We define the total correlation of the foreground as $\sum_{l,ij}\mathbf{P}_{l,ij}^{\mathrm{s}}\mathbf{S}_{l,ij}^{\mathrm{s}}$ and the total correlation of the background as $\sum_{l,ij}\mathbf{P}_{l,ij}^{\mathrm{b}}\mathbf{S}_{l,ij}^{\mathrm{b}}$, and then obtain the optimal foreground correlation tensors $\mathbf{P}_l^{\mathrm{s}}$ and the optimal background correlation tensors $\mathbf{P}_l^{\mathrm{b}}$ by maximizing the two total correlations. Each possible correspondence should have a confidence value, and we define the correlation tensor as $\mathbf{P}_l^{\mathrm{s}} \in [0,1]^{M_p\times N_p}$ and $\mathbf{P}_l^{\mathrm{b}} \in [0,1]^{M_p\times N_p}$, $M_p = N_p = H_p \times W_p$. We formulate this as an entropy-regularized optimal transport problem and solve it using the Sinkhorn algorithm [23] (see Algorithm 1). We change the correlation maps ($\mathbf{S}_l^s$ and $\mathbf{S}_l^{\mathrm{b}}$) into cost matrices ($\mathbf{C}_l^{\mathrm{s}}$ and $\mathbf{C}_l^{\mathrm{b}}$) by flipping the sign

of each element of the correlation maps. Notably, OBFTNeT uses ReLU to force the negative similarity to zero to suppress noisy correlation matches.

Algorithm 1 Optimal transport with Sinkhorn algorithm.

Input: $\mathbf{C}_l^{\mathrm{b/s}}, \epsilon, t_{max}$(The maximum number of iterations)

Initialize $\mathbf{K} = e^{-\mathbf{C}_l^{\mathrm{b/s}}/\epsilon}, \mathbf{n} \leftarrow \mathbf{1}, t \leftarrow 0$

while $t \leq t_{max}$ and not converge **do**

  $\mathbf{m} = 1/(\mathbf{K}\mathbf{n})$

  $\mathbf{n} = 1/\left(\mathbf{K}^{\top}\mathbf{m}\right)$

end while

Output: $\mathbf{P}^{\mathrm{b/s}} = \mathrm{ReLU}(\mathrm{diag}(\mathbf{m})\mathbf{K}\,\mathrm{diag}(\mathbf{n}))$

To facilitate the 4D convolution computation, we reshape $\mathbf{P}_l^{\mathrm{s}} \in \mathbb{R}^{M_p \times N_p}$ to $\mathbf{P}_l^{\mathrm{s}} \in \mathbb{R}^{H_p \times W_p \times H_p \times W_p}$, and reshape $\mathbf{P}_l^{\mathrm{b}} \in \mathbb{R}^{M_p \times N_p}$ to $\mathbf{P}_l^{\mathrm{b}} \in \mathbb{R}^{H_p \times W_p \times H_p \times W_p}$. We collect $\mathbf{P}_l^{\mathrm{s}}$ and $\mathbf{P}_l^{\mathrm{b}}$ with the same spatial size and denote the subset as $\{\mathbf{P}_l^{\mathrm{s}}\}_{l=1}^{L}$ and $\{\mathbf{P}_l^{\mathrm{b}}\}_{l=1}^{L}$, respectively. All optimal correlation tensors in $\{\mathbf{P}_l^{\mathrm{s}}\}_{l=1}^{L}$ and $\{\mathbf{P}_l^{\mathrm{b}}\}_{l=1}^{L}$ are concatenated along channel domain to form $\mathbf{P}_p^{\mathrm{s}} \in \mathbb{R}^{L \times H_p \times W_p \times H_p \times W_p}$ and $\mathbf{P}_p^{\mathrm{b}} \in \mathbb{R}^{L \times H_p \times W_p \times H_p \times W_p}$, where L is the number of CNN layers at pyramid layer p.

Finally, by concatenating (Cat) the optimal foreground correlation tensors $\mathbf{P}_p^{\mathrm{s}}$ and the optimal background correlation tensors $\mathbf{P}_p^{\mathrm{b}}$ of identical spatial size along channel domain at each pyramidal layer p to obtain a set of pyramid optimal bilateral correlation tensors $\{\mathbf{P}_p^{\mathrm{bi}}\}_{p=1}^{P}$ as

$$\{\mathbf{P}_p^{\mathrm{bi}}\}_{p=1}^{P} = \{\mathrm{Cat}(\mathbf{P}_p^{\mathrm{s}}, \mathbf{P}_p^{\mathrm{b}})\}_{p=1}^{P} \tag{4}$$

D. Correlation Analysis Module

Inspired by [33], we compress the $\{\mathbf{P}_p^{\mathrm{bi}}\}_{p=1}^{P}$ in the guidance dimension through the correlation analysis module to compress it into a 2D feature tensor, as shown in Fig. 4.

1) Inter-source bilateral correlation tensor encoder: At each pyramid layer p, a 4D convolution block compresses the guidance dimensions of $\mathbf{P}_p^{\mathrm{bi}}$ to $(H_\varepsilon, W_\varepsilon)$, while the query dimensions remain the same as (H_p, W_p) to achieve the transform from guidance information to query information, 4D convolution [33], [37], group normalization [38], and ReLU activation as illustrated in Fig. 5(a)-(c).

2) Inter-scale bilateral correlation tensor fusion: For the compressed multi-scale bilateral correlation tensors, we first transfer the fine bilateral correlation tensor with rich location information to the coarse bilateral correlation tensors from the bottom up (see Fig. 5(d)), and then transfer the bilateral correlation tensor with high-level semantic information to a lower level bilateral correlation tensor with rich spatial information transferring from the top down (see Fig. 5(e)). The bidirectional fusion process helps to establish a hierarchical relationship between different bilateral correlation tensors. Finally, 2D feature maps $\mathbf{P}_{p=1}^{7} \in \mathbb{R}^{128 \times H_1 \times W_1}$ are generated by compressing the bilateral correlation tensor through average-pooling in the guidance dimensions, where H_1 and W_1 are the height and width of the feature maps at pyramidal layer 1, respectively.

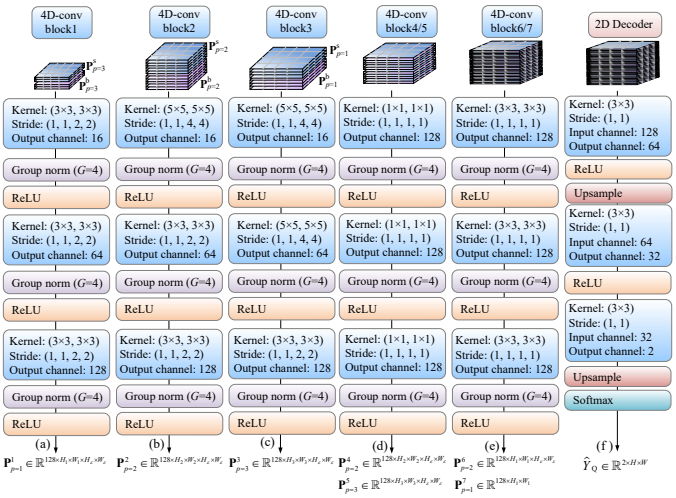

Fig. 5. 4D Convolution Block Details.

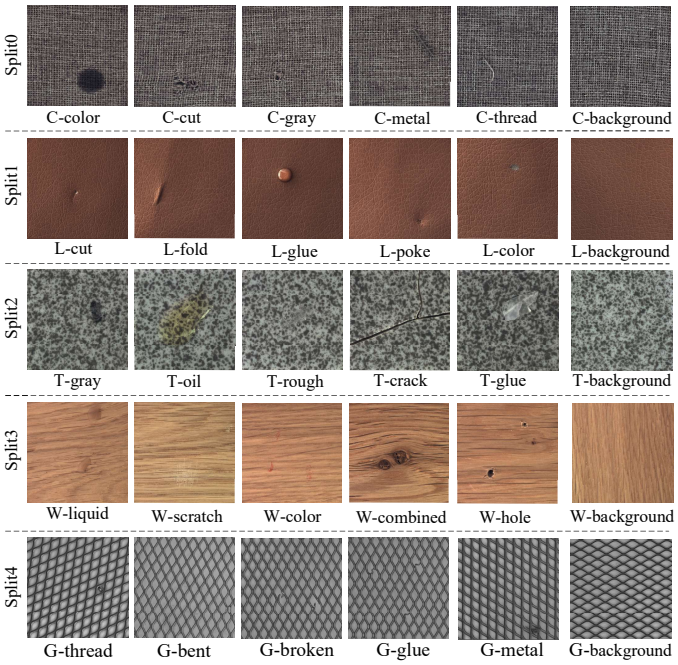

Fig. 6. Examples of the defect and defect-free samples.

3) 2D convolutional decoder: The 2D convolutional decoder as illustrated in Fig. 5(f), is composed of a series of stacked 2D convolutions, upsampling layers, and softmax function to obtain two channel map as $\hat{Y}_{\mathrm{Q}} = \text{2D-Decoder}\left(\mathbf{P}_{p=1}^{7}\right) \in [0,1]^{2 \times H \times W}$. During training, OBFTNet optimizes the model parameters using the mean of the cross-entropy loss of $\hat{Y}_{\mathrm{Q}}$ and ground-truth Y_{Q} overall pixel locations. During testing, we take the maximum channel value per pixel to get the query mask prediction $\overline{Y}_{\mathrm{Q}} \in \{0,1\}^{H \times W}$, where H and W are the height and width of the query mask prediction, respectively.

IV. Experiment

TABLE I
STATISTICAL OVERVIEW OF THE MVTEC AD DATASET (TEXTURES DEFECT)

Split	Material	Category and quantity					Background
0	Carpet (C)	color (19)	cut (17)	hole (17)	metal contamination (17)	thread (19)	background (280)
1	Leather (L)	cut (19)	fold (17)	glue (19)	poke (18)	color (19)	background (245)
2	Tile (T)	gray stroke (16)	oil (18)	rough (15)	crack (17)	glue strip (18)	background (230)
3	Wood (W)	liquid (10)	scratch (21)	color (8)	combined (11)	hole (10)	background (247)
4	Grid (G)	thread (11)	bent (12)	broken (12)	glue (11)	metal contamination (11)	background (264)

A. Defect Dataset

1) MVTec-Unseen: We use the real-world anomaly segmentation dataset MVTec AD [39] to evaluate our proposed approach. As can be seen from Fig. 6, the same industrial product has different types of defects, and the same defects also have differences in shape, size, color, and other characteristics. Specifically, industrial defects have the following characteristics: first, industrial defects are often non-salient as they are only a small area in high-resolution images (e.g., Fig. 6(C-gray)), and the defects have low contrast compared with the surrounding background (e.g., Fig. 6(T-rough)). Second, the complex background (e.g., Fig. 6(G-glue)) of many materials can also make the defect difficult to distinguish. Third, there is a large number of defect-free samples (column 6) in industry, while defect samples are limited.

To verify the practicality and effectiveness of our proposed method, we further partition the MVTec AD dataset. The 25 different categories of texture defects are divided into five splits according to the material type, and the materials of each split are the same, which is called MVTec-Unseen. The types and quantities of defects for different materials are shown in I. When testing the model on one split, we use the other four splits to train the model for cross-validation. This setting is practical in real-world applications as new defect segmentation tasks usually involve new classes and new domains (materials). Therefore, this setting is to test the generalization of the OBFTNet to unseen domains with different data distributions.
2) NEU-Seg: The NEU-Seg dataset [3] is used for the defect recognition problem of the hot-rolled strip, which contains three categories (inclusion, patch, and scratches). Each category contains 300 images and corresponding masks, and the resolution of each original image is 200×200. The biggest feature of the samples in NEU-Seg is the low contrast between the defect and the surrounding background area, and there is a large domain gap with the samples in the MVTec-Unseen dataset.

B. Implementation Details and Evaluation Metric

1) Implementation Details: Our model is implemented using PyTorch and runs on a server with an Nvidia RTX 3090 GPU card and Intel Xeon Silver 4210R CPU @ 2.40GHz. We choose VGG16 [40], ResNet50 [41], and ResNet101 [41] as our backbone networks for fair comparison with other methods. For VGG16 backbone, we extract features after every conv layer in the last two building blocks: from conv4_x to conv5_x, and after the last maxpooling layer. For ResNet backbones, we extract features at the end of each bottleneck before ReLU activation: from conv3_x to conv5_x. This feature extraction scheme can generate 3 pyramid layers (P=3) for each backbone network. All backbone networks are initialized with ImageNet [42] pretrained weights. The original spatial size of the image in the MVTec-Unseen is 512 $\times$ 512, and the original spatial size of the image in the NEU-Seg is 200 $\times$ 200. During training, we set the spatial size of support, query, and background images to 400 × 400 and batch size to 8. During the evaluation, each input sample is resized to the training patch size. In the optimal bilateral feature transport algorithm, we set $t_{max} = 20$ and $\epsilon = 0.05$ to calculate the optimal bilateral correlation tensors for each set of images $(X_{\mathrm{S}}, X_{\mathrm{Q}}, X_{\mathrm{B}}, Y_{\mathrm{S}})$. The objective function is the cross-entropy loss. The network uses the Adam optimizer with an initial learning rate of 0.0001 for 500 epochs. Once the model is trained, the texture defect segmentation model parameters are fixed and do not require optimization for the testing set. For unseen material few-shot texture defect segmentation, we compare OBFTNet with the the state-of-the-art methods SG-one [34], PANet [15], CANet [17], PMMs [18], PFENet [16], ASGNet [19], HSNet [33], and VAT [32].

2) Evaluation Metric: Following few-shot semantic segmentation methods [11], [15]–[19], [21], [22], [34], [35], [43], [44], we adopt the intersection over union (IoU), mean intersection over union (mIoU), and foreground-background IoU (FB-IoU) as our evaluation metrics for all experiments. We define the Intersection over Union (IoU) of class g as $\mathrm{IoU}_g = \frac{TP_g}{TP_g+FP_g+FN_g}$, where the TP_g, FP_g, and FN_g are the number of true positives, false positives, and false negatives of the predicted masks respectively. The mIoU is the average of IoUs over different classes as $\mathrm{mIoU} = \frac{1}{G}\sum_{g=1}^{G}\mathrm{IoU}_g$, where G is the number of classes in each split. We mainly use the mIoU metric because it takes into account the differences between the foregrounds of different defect classes and thus more accurately reflects the model performance. FB-IoU computes the average of foreground and background IoU regardless of object class, which reflects how well the full object extent is activated. The formulation follows $\text{FB-IoU} = \frac{1}{2}\left(\mathrm{IoU}_{\mathrm{F}} + \mathrm{IoU}_{\mathrm{B}}\right)$, where $\mathrm{IoU}_{\mathrm{F}}$ is the foreground IoU values, $\mathrm{IoU}_{\mathrm{B}}$ is the background IoU values. For the FB-IoU calculation for each split, only the foreground and background are considered ($G = 2$).

Implementation speed is an important measure of model efficiency, and we use average frame-per-second (FPS) to evaluate the efficiency of our model and comparison mod-

TABLE II
PERFORMANCE OF ONE-SHOT / FIVE-SHOT SEMANTIC SEGMENTATION USING MVTEC-UNSEEN. BOLD NUMBERS INDICATE THE BEST RESULTS, AND UNDERLINED NUMBERS INDICATE THE SECOND BEST RESULTS

Backbone	Method	one-shot								five-shot							
		split0	split1	split2	split3	split4	mean IoU	mean FB-IoU	FPS	split0	split1	split2	split3	split4	mean IoU	mean FB-IoU	FPS
VGG16	SG-one (TCYB2020) [34]	35.89	23.58	12.79	15.82	2.14	18.04	52.42	35.7	37.25	24.24	12.80	16.02	2.10	18.48	55.31	21.7
	PANet (ICCV2019) [15]	18.59	25.28	39.75	26.11	3.26	22.60	55.96	56.1	35.96	15.19	47.60	39.03	3.11	28.18	59.29	16.1
	PMMs (ECCV2020) [18]	28.78	16.45	10.79	10.85	2.10	13.80	50.21	40.0	29.04	16.34	11.77	10.79	2.35	14.06	53.46	17.2
	PFENet (TPAMI2022) [16]	19.69	5.85	17.67	12.42	4.11	11.95	49.45	37.0	18.46	16.38	23.85	15.19	4.72	15.72	51.70	15.1
	HSNet (ICCV2021) [33]	19.76	14.65	25.70	24.84	5.48	18.01	57.62	34.4	37.92	21.96	33.94	31.80	9.33	26.99	62.55	6.2
	OBFTNet (ours)	**46.85**	**47.21**	**46.31**	**46.63**	**25.88**	**42.58**	**68.82**	27.0	**47.21**	**46.98**	**47.97**	**46.52**	**26.59**	**43.05**	**69.79**	3.9
ResNet50	CANet (CVPR2019) [17]	17.59	24.18	26.05	6.24	6.90	16.19	47.62	71.4	-	-	-	-	-	-	-	-
	PMMs (ECCV2020) [18]	2.38	1.20	11.59	5.63	1.29	4.42	44.71	28.6	2.37	1.28	11.56	5.63	1.26	4.42	44.92	13.2
	ASGNet (CVPR2021) [19]	4.64	1.31	33.73	7.54	1.80	9.80	29.49	20.8	2.73	1.15	28.40	8.06	2.47	8.56	35.33	8.6
	PFENet (TPAMI2022) [16]	3.94	23.87	34.58	24.4	4.12	18.18	50.61	18.9	10.09	37.72	32.00	21.59	6.70	21.62	54.51	7.0
	VAT (arXiv21) [32]	39.19	33.83	15.38	18.77	4.01	22.24	57.12	15.1	41.31	35.08	15.85	21.45	4.28	23.59	57.87	1.9
	HSNet (ICCV2021) [33]	23.77	20.09	32.97	43.15	2.94	24.58	57.21	17.5	31.53	36.42	31.43	44.56	6.20	30.03	62.84	3.8
	OBFTNet (ours)	**56.34**	**54.95**	**56.39**	**52.79**	**40.29**	**52.15**	**74.76**	15.2	**56.80**	**55.22**	**57.54**	**53.57**	**40.00**	**52.63**	**75.07**	3.2
ResNet101	ASGNet (CVPR2021) [19]	14.70	9.11	44.67	7.21	1.37	15.41	39.22	12.3	16.53	22.47	26.52	6.06	1.28	14.57	41.61	4.3
	PFENet (TPAMI2022) [16]	8.49	7.36	37.28	10.08	2.18	13.08	47.91	9.6	5.52	30.38	46.55	13.88	3.28	19.92	49.49	4.0
	VAT (arXiv21) [32]	42.87	13.41	14.45	31.07	2.29	20.82	54.80	10.3	45.61	14.40	14.42	34.04	2.74	22.24	55.82	1.4
	HSNet (ICCV2021) [33]	19.56	32.72	39.90	40.62	7.87	28.13	61.88	12.0	28.88	41.13	47.13	43.95	14.69	35.16	65.36	2.4
	OBFTNet (ours)	**55.24**	**59.20**	**65.52**	**62.32**	**30.88**	**54.63**	**77.24**	9.1	**55.36**	**59.61**	**67.27**	**63.46**	**31.08**	**55.36**	**78.56**	1.7

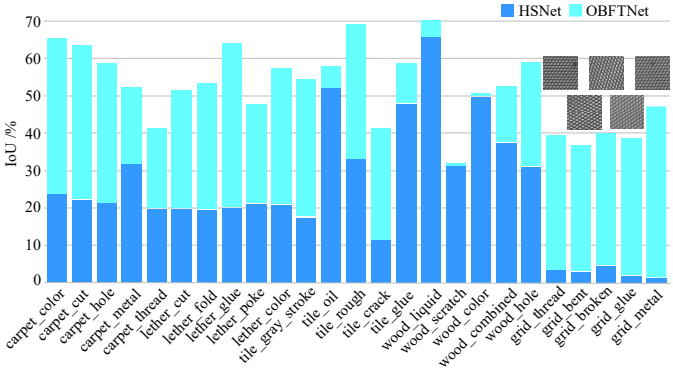

Fig. 7. Category-wised performance (IoU) gains on the MVTec-Unseen dataset. Our method (OBFTNet) achieves significant improvements compared with the baseline (HSNet).

els. During the evaluation, each input sample is resized to 400×400.

C. Cross-Domain Evaluation on MVTec-Unseen

In this experiment, a domain gap exists between the training and testing sets to facilitate the validation of the model's generalization to texture defects in unseen materials (domains). We compare our method with state-of-the-art methods on the MVTec-Unseen dataset. For the performance of the MVTec-Unseen dataset in the mIoU and FB-IoU metric, our method significantly outperforms state-of-the-art methods. As shown in Table II, using three different backbone networks, our model outperforms the state-of-the-art algorithms. With ResNet101 as the backbone network, our model achieves the best performance, achieving 26.50% (54.63% vs. 28.13%) and 20.20% (55.36% vs. 35.16%) of mIoU improvements over HSNet for one-shot and five-shot settings, respectively.

The improved performance of our proposed method is mainly attributed to the optimal bilateral feature transport strategy. It adaptively generates task-specific high-quality bilateral semantic guidance between guidance and query images conditioned on a small guidance set, overcoming domain gaps and background interference. Furthermore, we establish semantic correspondences between multi-scale guidance features and query features to deal with the problem of defect scale variation. However, conventional few-shot segmentation methods [15]–[19], [32]–[34] are designed for categories in the PASCAL VOC and MS COCO datasets [16], which have a large number of samples. The objects in these datasets are relatively large and have high inter-categories similarity. When these methods are applied to defect segmentation, due to the small defect size, serious background interference, and most importantly, the obvious domain gap between defects of different materials, this results in a significant decrease in segmentation performance. In particular, texture defects in the manufacturing industry often exhibit limited appearance information and few defect samples, which increases the difficulty of feature learning and results in many segmentation failures.

With the ResNet50 backbone network and one-shot setting,

the category-wised IoU of our method (OBFTNet) and the baseline method (HSNet) are shown in Fig. 8. Our method achieves significant improvements compared with the baseline method on all materials. The most noteworthy material for performance gains is the grid. This material has a complex background, and the contrast between the defect area and the surrounding background is low. OBFTNet has a larger performance gain on the defect category of this grid, showing its potential to handle background complex materials.

Previous methods are usually only able to deal with specific defects. Our method can segment various styles of defects from different materials, and our proposed OBFTNet performs well even in challenging situations, as evident in Fig. 7. For example, when the objects in the query image are small (rows 1-2), the defect has a low contrast to the surrounding background (rows 1-3), there are many objects in the query image (row 4), and the background is very complex (row 5), our OBFTNet can still produce segmentation results close to the ground truth. This further demonstrates the effectiveness and practicality of our proposed method for the task of industrial defect segmentation. However, the baseline method (HSNet) suffers from severely missed segmentation (rows 1, 3, 4, and 5) due to the above factors.

To verify the efficiency of our method, we test the FPS (average frame-per-second) of our method and the comparison methods, respectively. As shown in Table II, OBFTNet FPS is comparable to the other methods in terms of segmentation efficiency while the evaluation metrics (IoU, mIoU, and FB-IoU) are significantly better. Our proposed method can meet the usage requirements of different defect segmentation tasks by choosing different backbone networks and k-shot settings. For example, when the defect segmentation task requires high detection accuracy and speed, it is recommended to use the one-shot setting and the ResNet50 backbone network, which can achieve a segmentation speed of 15.2 FPS and a mIoU of 52.15%. When the defect segmentation task requires high segmentation speed, it is recommended to use the one-shot setting and the VGG16 backbone network, which can achieve a segmentation speed of 27.0 FPS and a mIoU of 42.58%. In conclusion, our proposed method can meet the usage requirements of different industrial defect segmentation tasks.

D. Cross-Domain Evaluation on NEU-Seg

We did not train our network using the NEU-Seg dataset for steel materials, but in order to verify the generalizability of our proposed method, we cross-domain the models trained on the MVTec-Unseen datasets to the NEU-Seg dataset to test the performance of OBFTNet on steel surface defects. In this experiment, we train a selection of few-shot segmentation models on MVTec-Unseen and then directly use them with the NEU-Seg dataset without any domain-specific model retraining or fine-tuning. This is a good cross-domain test, as both datasets exhibit clear domain shifts in the categories of instance size, number of instances, and image size. We choose the state-of-the-art method HSNet as the baseline. In Table III, with the ResNet50 backbone, our method has significant advantages in almost all splits. With the one-shot and five-shot settings,

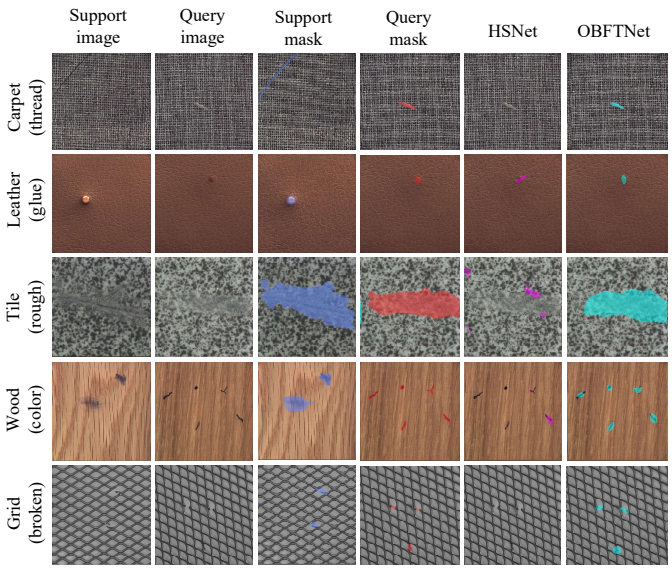

Fig. 8. Segmentation results of the proposed OBFTNet and the baseline (HSNet) for one-shot segmentation on MVTec-Unseen.

our proposed method achieves 10.41% (30.63% vs. 20.22%) and 13.32% (37.17% vs. 23.85%) performance (mIoU) improvement over the baseline, respectively. Furthermore, our performance advantage on FB-IoU is still obvious compared to the baseline method.

A visual comparison of the segmentation performance of our method and the baseline method (HSNet) on the NEU-Seg dataset is shown in Fig. 9. Achieving improvements on unseen (MVTev-Unseen) domains is more challenging due to the large gap between seen and unseen (NEU-Seg) domains, the low contrast between defect and background regions, and the scarcity of labeled samples for unseen tasks. We observed that in challenging defect segmentation cases such as low contrast (row 1), high intra-class variance (rows 3 and 5), and small objects (row 6), our proposed OBFTNet can still perform accurate segmentation. However, the baseline method has problems with false (rows 1-3) and missed (rows 4-6) segmentation, resulting in unreliable segmentation results. For defects with low intra-class variances and relatively high contrast (rows 2 and 4), OBFTNet can obtain more accurate segmentation results.

E. Ablation Studies and Discussion

We conduct ablation studies using the MVTec-Unseen dataset to analyze how each component affects the performance of the proposed method. We mainly consider four components: Optimal feature transport (OFT); optimal bilateral feature transport (OBFT); inter-scale bilateral correlation tensor fusion (BF); the number of iterations of the Sinkhorn algorithm. In all ablation studies, OBFTNet adopts ResNet50 as the backbone network.

1) Optimal feature transport: As described in Section III, the function of the optimal feature transport is to obtain a global optimal feature transport strategy from the support image to the query image, and it is a crucial part of

TABLE III
THE PERFORMANCE OF MVTEC-UNSEEN CROSS-DOMAIN TO NEU-SEG ON SEMANTIC SEGMENTATION UNDER ONE-SHOT/FIVE-SHOT SETTINGS

Backbone	Method	one-shot							five-shot						
		split0	split1	split2	split3	split4	mean IoU	mean FB-IoU	split0	split1	split2	split3	split4	mean IoU	mean FB-IoU
ResNet50	PMMs (ECCV2020) [18]	11.25	10.77	16.43	11.58	11.68	12.34	45.03	12.36	11.97	17.55	12.44	14.31	13.73	47.67
	HSNet (ICCV2021) [33]	23.27	20.48	15.60	19.93	21.83	20.22	56.25	25.17	26.10	17.60	23.40	26.97	23.85	58.29
	OBFTNet (ours)	**30.02**	**33.05**	**26.55**	**33.38**	**30.13**	**30.63**	**61.69**	**37.23**	**40.63**	**31.92**	**40.27**	**35.80**	**37.17**	**65.46**

TABLE IV
A ABLATION STUDY OF ONE-SHOT/FIVE-SHOT

OFT	OBFT	BF	one-shot							five-shot						
			split0	split1	split2	split3	split4	mean IoU	mean FB-IoU	split0	split1	split2	split3	split4	mean IoU	mean FB-IoU
			23.77	23.00	33.49	43.57	3.67	25.50	58.31	32.15	37.63	31.82	45.31	7.17	30.82	64.15
✓			39.77	42.06	**57.89**	39.27	6.87	37.17	67.51	50.49	51.20	**60.35**	44.12	28.83	47.00	72.32
✓	✓		54.54	54.30	55.07	50.14	39.36	50.68	73.18	55.04	**55.71**	56.92	50.46	**40.48**	51.72	74.24
✓	✓	✓	**56.34**	**54.95**	56.39	**52.79**	**40.29**	**52.15**	**74.76**	**56.80**	55.22	57.54	**53.57**	40.00	**52.63**	**75.07**

TABLE V
PERFORMANCE BASED ON NUMBER OF ITERATIONS OF THE SINKHORN ALGORITHM

t_{max}	split0	split1	split2	split3	split4	mean IoU	mean FB-IoU
5	51.46	52.94	55.17	**57.47**	39.53	51.31	73.39
10	56.76	54.36	55.73	53.06	40.35	52.05	74.15
20	56.34	**54.95**	**56.39**	52.79	40.29	**52.15**	**74.76**
30	56.70	54.42	56.01	52.78	**40.48**	52.08	74.53
40	**56.87**	54.75	55.70	53.05	39.70	52.01	74.45

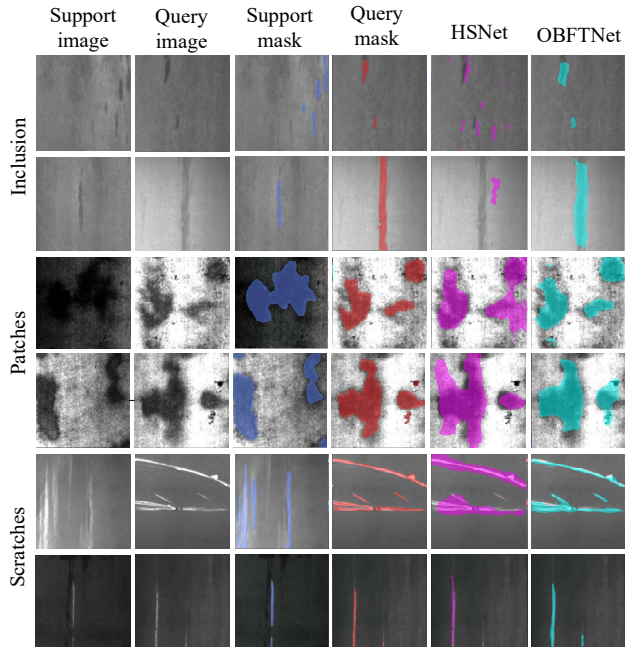

Fig. 9. Segmentation results on one-shot segmentation of the proposed OBFTNet and the baseline (HSNet) using a model trained on the MVTec-Unseen dataset cross-domain to the NEU-Seg dataset.

OBFTNet. To quantitatively analyze it, we use the MVTec-Unseen dataset to compare networks with and without optimal feature transport. As shown in Table IV, when the model uses optimal feature transport to obtain a global optimal feature transport strategy from the support image to the query image, the performance increases by 11.67% (37.17% vs. 25.50%) for one-shot and 16.18% (47.00% vs. 30.82%) for five-shot because optimal feature transport reduces false matching and background information interference.

2) Optimal bilateral feature transport: To quantitatively analyze our approach, we use the MVTec-Unseen dataset to compare the network with and without the background image learning branch. As shown in Table IV, when we introduce the background image to reverse prediction, the model performance is improved by 13.51% (50.68% vs. 37.17%) for one-shot and 4.72% (51.72% vs. 47.00%) for five-shot. In particular, when the optimal background feature transport is introduced, the texture defect segmentation accuracy for the grid material is significantly improved. The reason is that the background of the grid material defect is very complex, and the foreground defect is not obvious. It is difficult to complete such a complex defect segmentation only by relying on the foreground information in the support image. The optimal bilateral correlation tensor can improve the robustness of the model to complex background noise and improve the discriminative ability of the model by establishing accurate bilateral semantic correspondence. As shown in Fig. 10($\mathbf{P}_l^{\mathrm{s}}$), more key position correspondences between query and support samples are established in the optimal foreground transport solver to propagate foreground information. As shown in Fig. 10($\mathbf{P}_l^{\mathrm{b}}$), the optimal background transport solver will obtain some minor weights from some pixels with high texture similarity in the background image to enhance the corresponding pixels of the query image. Accurate bilateral semantic correspondence improves the discriminative ability and domain adaptability of the model.

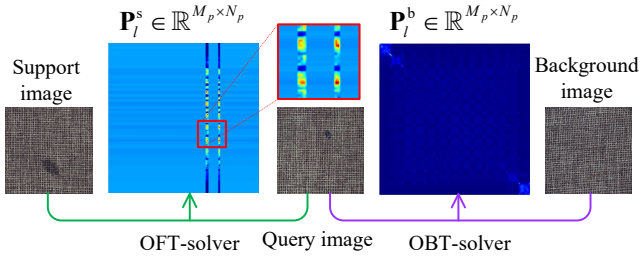

Fig. 10. Optima bilateral correlation tensors.

3) Inter-scale bilateral correlation tensor fusion: As described in Section III, we perform a bidirectional fusion of multi-scale bilateral correlation tensors. The reason is that most defects exhibit limited appearance information, and bidirectional fusion can effectively alleviate the missing segmentation problem. As shown in Table IV, the model performance is improved by 1.47% (52.15% vs. 50.68%) for one-shot and 0.91% (52.63% vs. 51.72%) for five-shot after using inter-scale bilateral correlation tensor fusion, which proves the effectiveness of bidirectional fusion.

4) Number of iterations of the Sinkhorn algorithm: In Table V, an ablation study is carried out to determine the impact of varying the number of iterations of the Sinkhorn algorithm. When t_{max} = 10 it significantly outperforms t_{max} = 5, which validates the plausibility of using optimal transport. When t_{max} = 20, the best performance is obtained. However, when t_{max} = 30 or 40, the performance gain is marginal. This proves that we can achieve a large performance gain with only a small number of iterations of the Sinkhorn algorithm.

V. Conclusion

Because of the problem that the current defect segmentation algorithm relies on large-scale data sets and the generalization ability of unseen material defects is poor, we propose an optimal bilateral feature transport network (OBFTNet) for few-shot texture defect segmentation. OBFTNet can adaptively establish a robust correspondence between the guidance image and the query image, maximizing the use of the available information. When used to segment unseen material defects, task-specific semantic correspondences can be adaptively generated conditioned on a small guidance set, overcoming domain gaps and texture shift issues. Our proposed method can shorten the application cycle of deep learning-based visual defect detection algorithms and reduce the model's dependence on large-scale datasets.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] 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.

[3] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7448–7458, 2019.

[4] J. Cao, G. Yang, and X. Yang, "A pixel-level segmentation convolutional neural network based on deep feature fusion for surface defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[5] G. Song, K. Song, and Y. Yan, "Edrnet: Encoder–decoder residual network for salient object detection of strip steel surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9709–9719, 2020.

[6] X. Ni, Z. Ma, J. Liu, B. Shi, and H. Liu, "Attention network for rail surface defect detection via consistency of intersection-over-union (iou)-guided center-point estimation," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1694–1705, 2021.

[7] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1266–1277, 2018.

[8] H. Yang, Y. Chen, K. Song, and Z. Yin, "Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 3, pp. 1450–1467, 2019.

[9] W. Liu, Z. Liu, H. Wang, and Z. Han, "An automated defect detection approach for catenary rod-insulator textured surfaces using unsupervised learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 8411–8423, 2020.

[10] H. Dong, K. Song, Q. Wang, Y. Yan, and P. Jiang, "Deep metric learning-based for multi-target few-shot pavement distress classification," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1801–1810, 2022.

[11] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[12] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5320–5329.

[13] R. Feng, X. Zheng, T. Gao, J. Chen, W. Wang, D. Z. Chen, and J. Wu, "Interactive few-shot learning: Limited supervision, better medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2575–2588, 2021.

[14] Y. Bao, K. Song, J. Liu, Y. Wang, Y. Yan, H. Yu, and X. Li, "Triplet-graph reasoning network for few-shot metal generic surface defect segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[15] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9197–9206.

[16] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, 2022.

[17] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[18] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proceedings of the European conference on computer vision*. Springer, 2020, pp. 763–778.

[19] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8334–8343.

[20] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.

[21] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9587–9595.

[22] B. Liu, J. Jiao, and Q. Ye, "Harmonic feature activation for few-shot semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3142–3153, 2021.

[23] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," *Center for Research in Economics and Statistics Working Papers*, no. 2017-86, 2017.

[24] H. Yang, Q. Zhou, K. Song, and Z. Yin, "An anomaly feature-editing-based adversarial network for texture defect visual inspection," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2220–2230, 2020.

[25] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[26] Y. Liu, L. Zhu, M. Yamada, and Y. Yang, "Semantic correspondence as an optimal transport problem," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4462–4471.

[27] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4937–4946.

[28] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao, "Unbalanced feature transport for exemplar-based image translation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 023–15 033.

[29] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "Mining latent classes for few-shot segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8701–8710.

[30] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," *CoRR*, vol. abs/1806.07373, 2018. [Online]. Available: http://arxiv.org/abs/1806.07373

[31] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," in *ICLR*, 2018.

[32] S. Hong, S. Cho, J. Nam, and S. Kim, "Cost aggregation is all you need for few-shot segmentation," *arXiv preprint arXiv:2112.11685*, 2021.

[33] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[34] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[35] X. Zhang, Y. Wei, Z. Li, C. Yan, and Y. Yang, "Rich embedding features for one-shot semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

[36] J. Min, J. Lee, J. Ponce, and M. Cho, "Learning to compose hypercolumns for visual correspondence," in *ECCV*, 2020.

[37] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," in *NeurIPS*, 2018.

[38] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[39] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] 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.

[43] H. Wang, X. Zhang, Y. Hu, Y. Yang, X. Cao, and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 730–746.

[44] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

Dexing Shan (Student Member, IEEE) received the B.S. and M.S. degree from the School of Mechanical Engineering and Automation, Liaoning Technical University, Fuxin, China, in 2017 and 2020, respectively. He is currently pursuing a Ph.D. degree in control engineering with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His current research interests include few-shot segmentation, defect segmentation, multi-modal fusion, and intelligent robot.

Yunzhou Zhang (Member, IEEE) received the B.S. and M.S. degrees in mechanical and electronic engineering from the National University of Defense Technology, Changsha, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009. He is currently a Full Professor with the College of Information Science and Engineering, Northeastern University. He also leads the Institute of Image Recognition and Machine Intelligence. His research interests include intelligent robot, computer vision, and sensor networks.

Sonya A. Coleman received the B.Sc. degree (Hons.) in mathematics, statistics, and computing and the Ph.D. degree in mathematics from Ulster University, Londonderry, U.K., in 1999 and 2003, respectively. She is currently a Professor with the School of Computing and Intelligent System, Ulster University, and a Cognitive Robotics Team Leader with the Intelligent Systems Research Centre. Her research has been supported by funding from various sources such as The Engineering and Physical Sciences Research Council (EPSRC), The Nuffield Foundation, The Leverhulme Trust, and the EU. She was involved in the EU FP7 funded projects RUBICON, VISUALISE, and SLANDAIL. She has authored or coauthored over 150 publications in robotics, image processing, and computational neuroscience.

Dr. Coleman was awarded the Distinguished Research Fellowship by Ulster University in recognition of her contribution to research in 2009.

Dermot Kerr received the B.Sc. degree (Hons.) in computing science and the Ph.D. degree in computing and engineering from Ulster University, Londonderry, U.K., in 2005 and 2008, respectively. He is currently a Lecturer with the School of Computing, Engineering and Intelligent System, Ulster University. He was involved in the EU FP7 funded projects VISUALISE and SLANDAIL. His current research interests include computational intelligence, biologically inspired image processing, mathematical image processing, omnidirectional vision, and robotics. Dr. Kerr is an Officer and a member of the Irish Pattern Recognition and Classification Society.

Dr. Kerr is also an Officer and a member of the Irish Pattern Recognition and Classification Society.

Shitong Liu received the bachelor's degree in automation from Qingdao University of Technology, Qingdao, China, in 2021. He is pursuing a master of engineering degree from the College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include few-shot learning, semantic segmentation, defect detection, and multi-modal fusion.

Ziqiang Hu received the B.S. degree in Mechanical Engineering and Automation from East China University Of Technology, Nanchang, China, in 2021. He is pursuing a master of engineering degree from the College of Information Science and Engineering, Northeastern University, Shenyang, China. His current research interests focus on few-shot learning, few-shot semantic segmentation, and self-supervised learning.