Industrial UAV-Based Unsupervised Domain Adaptive Crack Recognitions: From Database Towards Real-Site Infrastructural Inspections

Kangcheng Liu, *Member, IEEE*, and Ben M. Chen, *Fellow, IEEE*

***Abstract*—The defect diagnosis of modern infrastructures is crucial to public safety. In this work, we propose an unsupervised domain adaptive crack recognition framework. To fulfill the unsupervised domain adaptation (UDA) task of cracks recognition in infrastructural inspections, we propose a robust UDA learning strategy termed *Crack-DA* to increase the generalization capacity of the model in unseen test circumstances. More specifically, we first propose leveraging the self-supervised depth information to help the learning of semantics. And then using the edge information to suppress nonedge background objects and noises. We also use the data augmentation-based consistency. More importantly, we use the disparity in depth to evaluate the domain gap in semantics and explicitly consider the domain gap in network optimization. A database consisting of 11 298 crack images with detailed pixel-level labels for network training in domain adaptations is established. Extensive experiments on unmanned aerial vehicle (UAV)-captured highway cracks and real-site UAV inspections of building cracks demonstrate the robustness and effectiveness of the proposed domain adaptive crack recognition approach.**

***Index Terms*—3-D reconstructions, autonomous infrastructure inspections, crack detection and segmentation, domain adaptive learning, unmanned aerial vehicles (UAVs).**

I. Introduction

THE health condition monitoring of various infrastructures is of great significance to public safety [1], [2], [3]. Regular inspections of infrastructural health conditions in an autonomous way are vital to the subsequent repair and maintenance. From a technical perspective, inspections of infrastructures such as buildings and pavements require 3-D reconstruction methods of inspected targets to construct an accurate 3-D model, and defects recognition methods to realize the automatic defects identification [4]. With the development of robotics, autonomous systems such as unmanned aerial vehicles (UAVs), and unmanned ground vehicles (UGVs) have great potential to substitute humans to conduct cost-consuming and labor-intensive inspections [3]. Also, the learning-based methods can be applied in a data-driven manner to fault diagnosis [5], industrial anomaly detection [6], [7], and to accurately segment and localize defects such as cracks in captured images by autonomous UAV systems [8], [9], [10] in our applications of infrastructure inspections.

When deployed into real applications, several major challenges hinder the deployment of current learning-based crack recognition. First, an integrated autonomous system requires to be developed to fulfill the inspection task without human interventions. Specifically, advanced control and motion planning systems should be established to make the UAV conduct inspection tasks autonomously, and 3-D reconstruction methods should be developed to establish a 3-D model of the target infrastructure. Second, high-quality pixel-level labels of the inspection target are not always available. Therefore, we require the algorithm to perform well on the unlabeled test data with great domain adaptation capability. The third is the scarcity of 3-D structural information. Structural information has been demonstrated to be very useful in visual computing and recognition [11], [12]. However, the typical crack recognition methods [9], [10], [13], [14] have not incorporated the geometric information such as depth and edge into feature learning, which can not fully exploit useful features apart from semantics to further increase the discrimination capacity of the model. Finally, domain-invariant relationships between the semantics and geometric depth have not been explored to boost the performance in recognition of defects without labels when encountered with domain shifts, which limits the domain generalization capacity in real-site crack inspections.

To tackle the challenges above, we propose a domain adaptive learning framework *Crack-DA* with applications to industrial UAV-based crack inspections. Our proposed domain adaptive learning system can realize unsupervised domain adaptive recognitions of the defects such as cracks. It has been proved to be very effective for crack recognition in the unlabeled test images with large domain gaps when conducting a real-site building inspection. As depicted in Fig. 1, we have proposed

Manuscript received 11 April 2022; revised 3 July 2022 and 20 August 2022; accepted 23 August 2022. Date of publication 22 September 2022; date of current version 3 April 2023. This work was supported in part by the Hong Kong Centre for Logistics Robotics (HKCLR), and in part by Hong Kong PhD Fellowship Scheme. *(Corresponding author: Kangcheng Liu.)*

The authors are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong,, Hong Kong (e-mail: kcliu@link.cuhk.edu.hk; bmchen@cuhk.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2022.3204953.

Digital Object Identifier 10.1109/TIE.2022.3204953

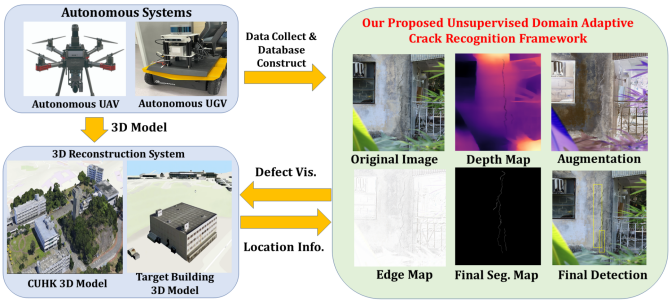

Fig. 1. We have proposed our unsupervised domain adaptive crack recognition framework highlighted by red font, termed *Crack-DA*. Integrated with autonomous systems to collect data and 3-D reconstruction system to build target 3-D model, our proposed framework can fulfill the automatic infrastructural inspection.

our own unsupervised domain adaptive crack recognition framework, termed *Crack-DA*, highlighted in red. Integrated with autonomous systems to collect data and 3-D reconstruction system to build target 3-D model, *Crack-DA* can fulfill the automatic infrastructural inspection without human labor. The contribution of our work is as follows:

1) We constructed a benchmark pixel-level crack segmentation database and open-sourced it for the community.
2) In addition, we propose a systematic framework termed *Crack-DA* to tackle the UDA of the crack recognition. The depth and edge information are fully exploited in proposed optimization functions and fusion networks are designed to fuse the depth with the semantic feature.
3) Furthermore, we propose using the disparity of depth estimation to evaluate the domain gap between the source and target domain. The estimated domain gap is formulated into the network optimization.
4) It is demonstrated that *Crack-DA* can greatly improve the model generalization capacity. Finally, real-site inspections of the highway pavement cracks, and building cracks demonstrate the robustness of the proposed system. *Crack-DA* is deployed to real-site inspections of cracks on pavements and buildings with good performance.

II. Related Work

Crack Recognition is a vehement research topic in last years. For establishing crack datasets, many recent studies addressed this issue by either manually collecting and labeling defect images [15] or synthesizing defect images [16]. Fully supervised learning-based methods have achieved great success in crack classification [17] and segmentation [13] when the domain gap between the training set and test set is not that large. Several pioneer works have used deep-network-based methods for fault identification [18], [19] and meta-learning-based methods for multiple-target defects recognition [17]. However, in real applications, the domain gap is very large when deploying the trained crack recognition model to the real-site target infrastructure to be inspected. According to our experiments, the performance of fully supervised crack recognition methods on the unseen tested target without labels is limited. Although we can utilize the autonomous UAVs to replace human power to conduct the inspection and capture the images of the target infrastructure, pixel-labeling for the images of the inspection target is time-and-labor-consuming. Therefore, a domain adaptation method that can leverage both labeled source domain data and unlabeled target domain data to boost the target-domain recognition performance is in great demand, and this makes our proposed work meaningful compared with related works in crack recognition. Also, the 3-D visions techniques have undergone tremendous progresses in the past few decades [20], [21]. However, unlike the 3-D data, which has small domain gaps in real circumstances, the 2-D images may have small domain gaps.

Unsupervised Domain Adaptation (UDA) denotes the model adaptation from the source domain with labels to the target domain without labels [22], [23]. The target of UDA is to eliminate the domain gap [24], [25], [26]. To achieve UDA, several recent studies tackle this challenge via image translation in the input space [27], adversarial learning in the feature space, or self-training and consistency learning in output space [28], [29], [30]. The first mainstream UDA approach is learning the domain-invariant feature distributions, and the second is conducting recursive self-training using highly confident network predictions as pseudolabels. In our implementation, we also use the self-training methods to continuously refine the pseudolabels on the target domain. Some multitask learning frameworks have been proposed recently, which apply a joint learning strategy for the semantic segmentation with instance segmentation [31] and boundary prediction [32]. However, these methods work in a fullysupervised setting, where domain gaps do not exist. It has been demonstrated that auxiliary visual representations such as the depth and edge information have strong correlations with the semantics of visual contents [33], [34], [35]. But the relationship between low-level geometry such as the depth and high-level semantics such as the pixel-level category has rarely been explored in the UDA setting with large domain gaps. Also, the exploration of UDA for specific defects recognition with a large domain gap is still in its infancy but vital and in great demand.

Autonomous Visual Systems for Inspections are of great significance to building a smart city. Several UAV-based datasets for generic object detection have been established, such as the MOR-UAV [36] for moving object detection in capture videos. The attention-based network for simultaneous object detection and counting has been proposed for UAV captured images [37]. However, two major problems exist. The first is that there exists no systematic framework for infrastructural inspections. The second is that current proposed datasets or methods are not specifically designed for unsupervised domain adaptive defects recognition. Therefore, a systematic framework dealing well with the domain gap for inspections with autonomous inspection systems requires to be constructed.

III. Methodology

As shown in Fig. 2, we have proposed our own unsupervised domain adaptive crack recognition framework, termed *Crack-DA*, highlighted by red font. Integrated with autonomous systems to collect data and 3-D reconstruction system to build target

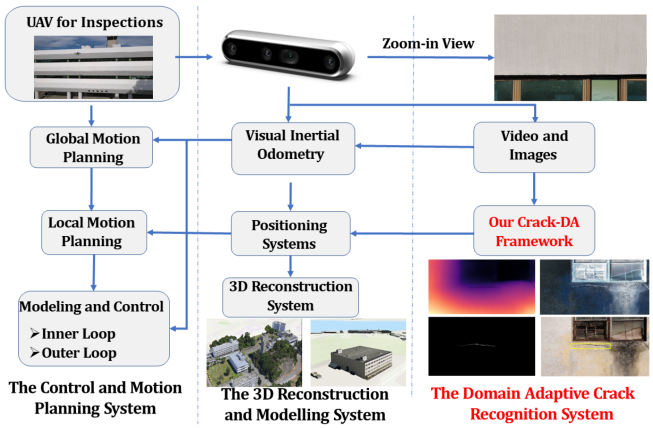

Fig. 2. Structure of the proposed autonomous inspection system. Our proposed domain adaptive crack recognition system is highlighted in red.

Fig. 3. CUHK United College Wu Chung Library final 3-D reconstruction results based on RGB images and the final infrared 3-D reconstruction results based on infrared images.

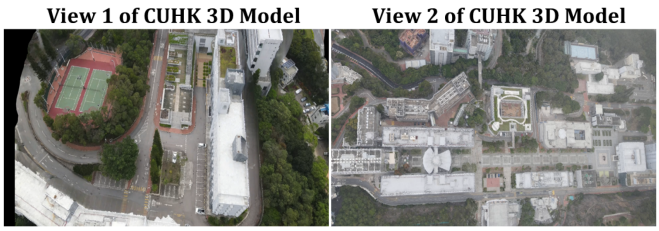

Fig. 4. Final 3-D reconstruction results of the CUHK campus. We randomly select two views for visualizations.

3-D model, our proposed framework can fulfill the automatic infrastructural inspection without intensive human labor.

We build our 3-D reconstruction system based on the SOTAs Open-MVG [38]. Apart from Open-MVG, the 3-D point clouds semantic segmentation algorithms [11] have been integrated for the building components segmentation. Our system supports versatile postprocessing functionalities. It can perform mesh generation and denoising with our texture mapping and denoising algorithms. Finally, we can render a structured 3-D mesh model with fine details. UAV-based reconstruction result examples of the CUHK library and campus are shown in Figs. 3 and 4. Our proposed reconstruction system can reconstruct a perceptually very accurate 3-D model, supporting both the RGB and infrared images as the input. The final detected defects are easily registered to the 3-D model using the 2-D–3-D correspondence.

A. Database Construction

We construct a comprehensive crack segmentation database consisting of 11 298 450×450 cracks images with detailed pixel-level labels.[1] The database contains cracks from various structures, including pavements, bridges, and buildings. In the formulation of domain adaptive crack recognition, we regard the proposed database as the source dataset with label and transfer the model to the pavement-crack and MaWan-crack unlabeled test set for crack recognition with proposed *Crack-DA*.

B. Unsupervised Domain Adaptive Crack Recognition

Take the building inspection as an example, after we have set up our database, it is still difficult to directly apply those data to practical infrastructural inspections. Because cracks from various infrastructures have intrinsically different patterns in geometric structures and background materials, directly testing trained models on target infrastructures with domain gaps inevitably causes a performance decrease. In this work, we propose a novel UDA framework termed *Crack-DA* to overcome the difficulties in crack recognition. Optimization functions and fusion networks are proposed for UDA.

1) Problem Definition: The target of UDA is transferring models trained on the source-domain labeled data to the target-domain unlabeled data, which matches our target of transferring models trained from the self-established database with labels to the target scenarios in inspections without labels. Provided the source domain images $X^{\mathbb{S}}$ in our self-established database with their corresponding labels $L^{\mathbb{S}}$, and the target domain crack images $X^{\mathbb{T}}$ captured by UAV without labels, the goal of the unsupervised domain adaptive crack segmentation is to learn a robust model $\mathbf{A}_{\text{seg}}$ that gives precise crack segmentation prediction in the target domain for inspected infrastructures. Denote the source domain images as $X^{\mathbb{S}} = \{(\mathbf{x}_1^S, \mathbf{l}_1^S, \mathbf{d}_1^S), (\mathbf{x}_2^S, \mathbf{l}_2^S, \mathbf{d}_2^S), \ldots, (\mathbf{x}_M^S, \mathbf{l}_M^S, \mathbf{d}_M^S)\}$, where $\mathbf{x}_i^S$ denotes the i_{th} training image, $\mathbf{l}_i^S$ denotes the i_{th} corresponding label for semantic segmentation, and $\mathbf{d}_i^S$ denotes the pseudoground truth in the auxiliary depth estimation task. Denote the target domain images as $X^{\mathbb{T}} = \{(\mathbf{x}_1^T, \mathbf{d}_1^T), (\mathbf{x}_2^T, \mathbf{d}_2^T), \ldots, (\mathbf{x}_M^T, \mathbf{d}_M^T)\}$, where $\mathbf{x}_j^T$ denotes the j_{th} real-scene test image. $\mathbf{d}_j^T$ denotes the corresponding pseudo label in the auxiliary depth estimation task.

For depth estimation, we use the off-the-shelf depth transformer [39] trained on KITTI [40], [41] to obtain the pseudodepth estimation ground truth on both the source and target dataset. The auxiliary depth information helps the network estimate the domain gap between the source and target domain, which facilitates optimizations in domain transfer.

2) Main Contributions: To tackle the great challenge that large domain gaps exist for the UDA in our applications, in this work, we propose effective network modules to improve the domain adaptation capacity. *First*, we propose using auxiliary geometric information from the depth estimation to achieve depth awareness. *Second*, we propose using an off-the-shelf edge

[1]https://github.com/KangchengLiu/Crack-Detection-and-Segmentation-Dataset-for-UAV-InspectionOur Self-Established Database

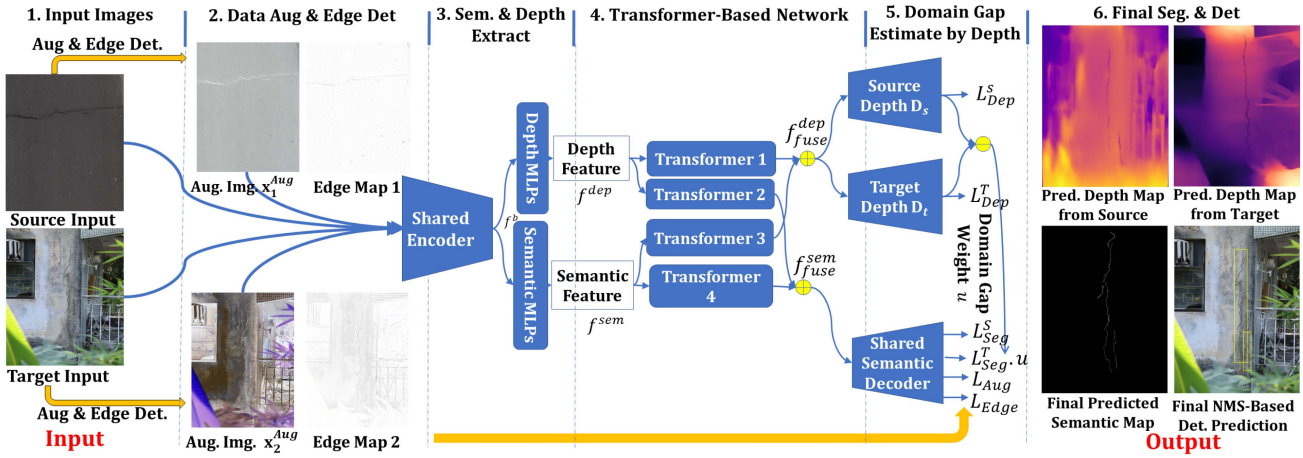

Fig. 5. Our proposed network framework *Crack-DA* for unsupervised domain adaptive crack segmentation. Our framework takes advantage of both depth and edge information for a more robust and accurate domain adaptive feature learning.

detector [42] to achieve edge awareness in the task of semantic segmentation. *Third*, we propose using the data augmentation approach to achieve consistencies in the feature learning and improve the robustness of the network prediction under various input data transformations. *Finally*, to improve the model adaptability to the target task, we propose using the disparity in depth prediction to evaluate the domain gap between the source and target domain. Then the domain gap is considered and formulated into the network's optimization functions to refine the semantic prediction of crack segmentation.

3) Crack-DA Network Architecture With Depth and Edge Awareness: As shown in Fig. 5, we propose a network with depth and edge awareness to achieve domain adaptive crack segmentation. The network consists of the following sequential steps to do semantic segmentation: 1) We obtain crack images of both source databases with pixel-level labels and target buildings captured by the UAV without labels. 2) We perform the data augmentation with wavelet transform [43] and inverted color and do Canny-based [42] edge detection for both the source domain images and the target domain images. 3) We use the shared semantic segmentation backbone encoder network to obtain the embedding depth feature and semantics feature, respectively. 4) We propose a transformer-based network for the relationship mining and fusing of depths and semantics. And we propose specifically designed optimization functions to achieve depth and edge awareness in learning. 5) Finally, we obtain the final segmentation and depth predictions. 6) Also, we apply selective search [44] based nonmaximum suppression (NMS) on the segmentation results to obtain the final crack detection results.

Transformer-Based Depth and Semantics Feature Fusion Network: To capture the correlations between the depth and semantics, we elaborately design the feature fusion network to fuse the depth feature with the semantic feature. And the attentional transformer (A-T) is used to model the relationship between the semantic features and the depth features. Specifically, after feeding the input images to the shared encoder, we can obtain the backbone feature $f_b \in \mathbb{R}^{512\times 1}$. Then we apply two-layers *depth perceptrons (MLPs)* and *semantics MLPs,* respectively, to obtain the depth feature f^{dep} and semantics feature f^{sem}. For the two-layers MLPs, the weight matrix of the first layer is $W_a \in \mathbb{R}^{512\times 256}$, and the weight matrix of the second layer is $W_b \in \mathbb{R}^{256\times 128}$. Then, we can obtain the depth feature $f^{\text{dep}} \in \mathbb{R}^{128\times 1}$ and semantic feature $f^{\text{sem}} \in \mathbb{R}^{128\times 1}$, respectively. Last, we apply four A-Ts with different weight matrices W_1, W_2, W_3, W_4 to obtain both the fused depth feature $f_{\text{fuse}}^{\text{dep}}$ and the fused semantic feature $f_{\text{fuse}}^{\text{sem}}$ in an adaptive manner

$$f_{\text{fuse}}^{\text{dep}} = W_1 f^{\text{dep}} + W_2 f^{\text{sem}} \tag{1}$$

$$f_{\text{fuse}}^{\text{sem}} = W_3 f^{\text{dep}} + W_4 f^{\text{sem}}. \tag{2}$$

The proposed attentional transformer has three prominent merits. First, the A-T can automatically select and enhance the closely related depth and semantics features, and the irrelevant features are suppressed. Second, the complementary information of depth and semantics are implicitly modeled and learned by the A-T. Third, the A-T is also used to capture the feature relationships between the source and the target domain. For example, the crack patterns are shared in the source and the target domain. Finding correlations and common features between domains can facilitate the learning of general feature representations for crack patterns. According to our ablation studies, the proposed attentional transformer for feature-fusion is significant to the overall recognition performance.

Optimization Functions for Robust Network Predictions: To ensure the decision boundary of the network lies in low-density regions and improve the robustness of the network prediction under input transformations, we propose optimization functions to encourage high similarity levels between the predictions for the two augmented input images. As shown in Fig. 5, denote the two augmented images as $\mathbf{x}_1^{\text{Aug}}$ and $\mathbf{x}_2^{\text{Aug}}$, they are fed into the shared encoder to obtain the semantic features. Then, the semantic features are fed into the shared semantic decoders to obtain the semantic segmentation prediction confidence vectors $\mathbf{p}_{1,j}$ and $\mathbf{p}_{2,j}$, where j is the pixel index in the image. Then, the highest confidence vectors in $\mathbf{p}_{1,j}$ and $\mathbf{p}_{2,j}$ are regarded as one-hot pseudo label vectors $\mathbf{y}_{1,j}$ and $\mathbf{y}_{2,j}$. Denote $\mathcal{B}$ as the whole

training set of input images, we formulate the data augmentation loss L_{Aug}

$$L_{\text{Aug}} = \frac{1}{\|\mathcal{B}\|}\sum_{\mathbf{x}_i\in\mathcal{B}}\frac{1}{N_p}\sum_{j=1}^{N_p}(\mathbf{Div}_{\text{JS}}(\mathbf{p}_{2,j}\|\mathbf{y}_{1,j}) + \mathbf{Div}_{\text{JS}}(\mathbf{p}_{1,j}\|\mathbf{y}_{2,j})). \tag{3}$$

$N_p = w \times h$ is the number of pixels within the input images. w and h denote the width and height of the input images. Jensen–Shannon (JS) divergence is used to encourage the consistency between the two augmented network predictions.

Optimization Functions With Edge and Depth Awareness: To exploit the edge and the depth information to facilitate the domain adaptation, we have proposed depth-aware optimization functions and also used the edge information to guide the optimization of the segmentation network to find accurate nonedge regions. From the principle of the image-based crack segmentation and our observations, the crack can merely appear at the pixels with great directional change in the RGB images. These pixels also have a large gradient in the neighborhood. Therefore, we propose an edge-aware optimization function to encourage the network to have correct predictions at the background noncrack pixels with a small gradient. We first use the Canny edge detector [42] to obtain the edge map. The edge map gives a great indication of the noncrack pixels. To be more specific, if the edge value is lower than a certain small value γ in the Canny edge detector ($\gamma = 0.2\%$ in our case), we regard the pixels as the noncrack pixels. Denote the noncrack pixels as $\mathbf{y}_n$, we encourage network predictions to classify these pixels as noncracks to suppress the noise and noncrack patterns. Denote the network predictions at the noncrack pixels as $\mathbf{p}_n$, and the number of the non-crack pixels as N_{nc}, we formulated our segmentation optimization function L_{edge} as follows:

$$L_{\text{edge}} = \frac{1}{\|\mathcal{B}\|}\sum_{\mathbf{x}_i\in\mathcal{B}}\frac{1}{N_{nc}}\sum_{n=1}^{N_{nc}}\|\mathbf{p}_n - \mathbf{y}_n\|^2. \tag{4}$$

By the constraints of the proposed $\mathcal{L}_2$-based optimization function, auxiliary edge supervisions are added to suppress background noises and to help noncrack pixels to be precisely segmented.

As no depth ground truth is provided, first, we acquire the pseudo ground truth depth as self-supervision from the transformer-based depth estimator [39]. To realize the depth awareness, we use the Berhu loss [45]. The Berhu loss L_{Berhu} is formulated as follows:

$$L_{\text{Berhu}}(z) = \begin{cases} \|z\| & \|z\| \le a \\ \frac{z^2+a^2}{2a} & \|z\| > a. \end{cases} \tag{5}$$

The Berhu loss resembles $\mathcal{L}_1$ loss when the norm of input z is less than a, while it resembles $\mathcal{L}_2$ loss when the norm of input z is larger than a. And a is the depth threshold set to $\frac{1}{5}$ of the maximum depth difference in the pseudoground truth depth map. Finally, both the source and target dataset can be used to learn the depth from the pseudolabels, which improves the 3-D geometric depth awareness within diverse 2-D scenes. The total optimization function for the depth estimation L_{dep} is formulated as the sum of the loss for the source dataset L^S_{dep} and the loss for the target dataset L^T_{dep}

$$L_{\text{dep}} = L^S_{\text{dep}} + L^T_{dep} = \frac{1}{\|\mathcal{B}\|}\sum_{\mathbf{x}_i\in\mathcal{B}}\frac{1}{N_p}\sum_{i=1}^{N_p}(L_{\text{Berhu}}(\|\mathbf{d}^S_{\text{pred},i} - \mathbf{d}^S_i\|) + L_{Berhu}(\|\mathbf{d}^T_{\text{pred},i} - \mathbf{d}^T_i)\|). \tag{6}$$

Domain Gap Estimation by Depth: To consider the domain gap between the source and target dataset, we utilize the difference in depth estimation to evaluate the domain gap between the source and target dataset in semantic segmentation. To be more specific, as shown in Fig. 5, let a real-site tested image feature $f^{\text{dep}}_{\text{fuse}}$ captured for the inspected target as input, we calculate the depth prediction for the test images using source depth decoder $\mathbf{D}_s$ and target depth decoder $\mathbf{D}_t$, respectively. Then, we obtain the predicted depth map from the source decoder and the predicted depth map from the target decoder. Because the depth and semantics are inherently and implicitly correlated, the semantic domain gap G_{sem} between the source and target dataset can be relatively precisely estimated by the prediction inconsistency in depth estimation. The domain gap weight u that uses the domain gap to estimate the confidence of the pseudo label at the target dataset domain is proposed

$$G_{\text{sem}} = \|\mathbf{D}_s(f^{\text{dep}}_{\text{fuse}}) - \mathbf{D}_t(f^{\text{dep}}_{\text{fuse}})\|$$
$$u = \exp\left(-\frac{G_{\text{sem}}}{b_{\max}}\right) \tag{7}$$

where $b_{\max}$ is the maximum depth difference in the pseudoground truth depth map of the target dataset. And the weight $u \in (0, 1]$. For a bigger domain gap, we will assign smaller weights to the segmentation loss function to reduce the confidence level of the semantic pseudolabels in optimizations and vice versa. In this way, the domain gaps are considered, which guide the network optimization.

Overall Optimization Functions: As shown in Fig. 5, after applying attentional transformers to obtain the fused depth feature $f^{\text{dep}}_{\text{fuse}}$ and the fused semantic feature $f^{\text{sem}}_{\text{fuse}}$, we input $f^{\text{sem}}_{\text{fuse}}$ to the shared semantic decoder to obtain the final segmentation prediction [46]. Also, we use two independent depth decoders that are trained with the source depth loss L^S_{dep} and the target depth loss L^T_{dep}, respectively. Denote the segmentation losses for the source and target dataset as L^S_{Seg} and L^T_{Seg}, respectively, we choose to use the focal loss [47] to tackle the great class imbalance for the crack segmentation task

$$L_{\text{Seg}} = -\frac{1}{\|\mathcal{B}\|}\sum_{\mathbf{x}_i\in\mathcal{B}}\frac{1}{N_p}\sum_{i=1}^{N_p} w(1-h^{gt}_i)^{\alpha}\log(h_i) + w(h^{gt}_i)^{\alpha}\log(1-h_i) \tag{8}$$

where h_i represents the binary predicted crack map and h^{gt}_i denotes the ground truth (in L^S_{Seg}) or pseudoground truth (in

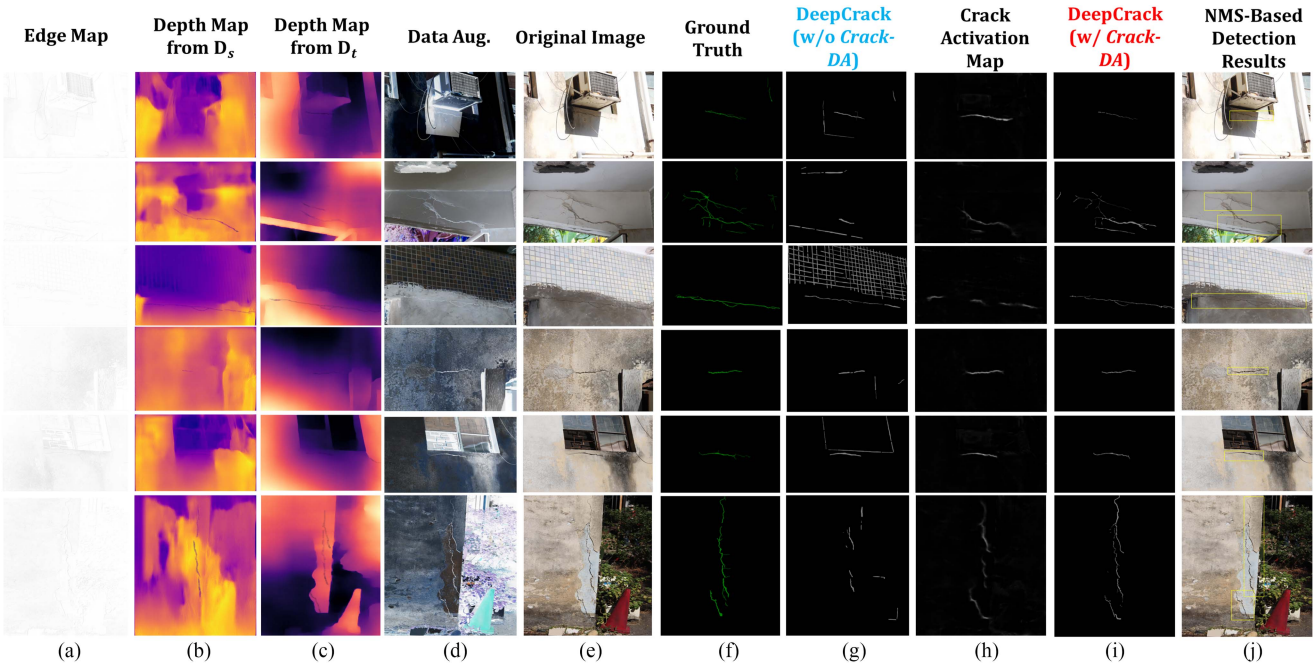

Fig. 6. Complete UAV-based real-site building cracks recognition qualitative results at MaWan, Hong Kong. The results without and with our proposed *crack-DA* are shown in blue and red, respectively.

L^T_{Seg}) crack map. We set the weight $w = 1$ in the segmentation loss L^S_{Seg} for the source dataset, and $w = u$ in the segmentation loss L^T_{Seg} for the target dataset. We select $\alpha = 2$ according to the original approach. N_p is the total number of pixels. Because the label in the target domain is not available for UDA, we use the self-training [48] approach which only believes the confident network prediction to be the pseudolabel with a confidence threshold σ. In summary, the total optimization function L_{total} for the *Crack-DA* training in an end-to-end manner to fulfill the UDA is formulated as

$$L_{\text{total}} = L_{\text{Aug}} + L_{\text{Edge}} + L_{\text{Dep}} + L^S_{\text{Seg}} + L^T_{\text{Seg}}. \tag{9}$$

IV. Experimental Results

A. Experiments of UDA From Our Self-Established Dataset to Real-Site Building Crack Inspections at MaWan, Hong Kong

To demonstrate the effectiveness of *Crack-DA*, we directly test it at MaWan, Hong Kong for the UAV-based inspections of cracks on old buildings.

Experimental Settings: We did experiments on our self-established dataset and transferred it to the MaWan old buildings' images captured by UAVs with the proposed *Crack-DA*. The confidence threshold σ is set to 0.8. The images in the source domain labeled dataset are in the resolution of 450×450. For the labeled data, we have used the whole source pixel-level labeled dataset consisting of 11 298 images for training. For the unlabeled target-domain data, we have used the 100 test set images with the resolution of 6000×4000 during the UAV building inspections at MaWan for testing.

Our proposed UDA framework *Crack-DA* is implemented with the deep learning framework *Pytorch*. We train our network with a single RTX 2070 GPU for 580 epochs with the Adam optimizer, utilizing a learning rate of 10^{-4}, which is multiplied by 0.1 per 60 epochs. Training takes approximately 16 h for DeepCrack [13] with *Crack-DA*. Note that all our proposed optimization function designs are merely required in the training stage and do not affect the efficiency. All our experimental results are three-time average.

Experimental Results: The qualitative experimental results for the MaWan building cracks are shown in Fig. 6. From column (a), we can see that our proposed method can provide an explicit edge map, which can clearly reveal the noncrack pixels. From columns (b) and (c), it can be seen that the domain gaps exist and the target decoder can offer more accurate depths and the patterns of cracks are effectively learned. In column (e), we show the original images captured for the building surface at MaWan by UAVs. While the data augmentation results with wavelet transform and inverted color are shown in column (d). It can be seen that the original image can be transformed into a different representation to enrich the training samples. We visualize the semantic segmentation results of previous SOTAs crack segmentation network DeepCrack [13] without domain adaptation as shown in the column (g) of Fig. 6. And the final segmentation predictions of DeepCrack [13] with our proposed *Crack-DA* are shown in column (i), which are far more accurate compared with the counterpart without domain adaptation. The various noncrack noises or objects in the background including bricks and grasses are successfully suppressed and eliminated. We also visualize the network activation map provided by our proposed optimization functions in column (h). It can be seen that our proposed network modules clearly help to find the

TABLE I
COMPARISON OF UDA SEGMENTATION RESULTS BETWEEN VARIOUS STATE-OF-THE-ART METHODS

Methods	Inference Times/ms	Global Accuracy (GA)/%	**MIoU**/%	Precision/%	Recall/%	F-Score/%
Crack-Net (w/o *Crack-DA*) [9]	98/465	51.3/49.6	43.5/41.2	41.3/39.2	40.5/38.8	40.9/39.0
Crack-Net-DA (w/ *Crack-DA*) [9]	106/488	79.2/75.6	72.5/70.3	71.9/70.2	72.5/70.3	72.2/70.3
DeepCrack (w/o *Crack-DA*) [13]	113/655	47.1/45.1	42.7/40.5	40.3/38.2	39.5/37.8	39.9/38.0
DeepCrack-DA (w/ *Crack-DA*) [13]	118/678	78.6/75.9	72.2/70.1	72.6/71.7	71.5/69.3	72.1/70.5
DeepLabV3 (w/o *Crack-DA*) [8]	101/466	48.2/45.3	42.7/39.2	49.3/47.5	48.9/47.8	49.1/47.7
DeepLabV3-DA (w/ *Crack-DA*) [8]	112/492	85.7/82.2	77.2/75.0	75.2/73.3	74.2/73.2	74.7/73.3
PSPNet (w/o *Crack-DA*) [52]	108/557	47.5/45.7	39.8/36.9	38.5/36.1	39.6/37.3	39.1/36.7
PSPNet-DA (w/ *Crack-DA*) [52]	126/579	83.4/80.2	73.2/70.8	72.9/69.8	73.6/70.7	73.2/70.2
ASPP-Net (w/o *Crack-DA*) [50]	122/745	49.9/47.5	44.6/42.7	46.6/44.2	46.7/44.9	46.7/44.5
ASPP-Net-DA (w/ *Crack-DA*) [50]	139/766	85.7/82.5	73.7/71.1	75.4/73.5	75.7/72.8	75.5/73.1
Seg-Former (w/o *Crack-DA*) [51]	132/787	55.3/52.6	43.3/41.1	47.2/43.7	44.2/42.9	45.7/43.3
Seg-Former-DA (w/ *Crack-DA*) [51]	151/808	86.2/83.5	78.9/76.2	78.8/74.3	78.9/75.2	78.9/74.7

"w/ *Crack-DA*": Results with the proposed UDA approach *Crack-DA*. "w/o *Crack-DA*": Results without *Crack-DA*. The left value of "/": Results for the highway cracks, and the right value of "/": Results for the real-site mawan building cracks. The network inference time is for a single test image. The MIoU is the most important metric in the crack segmentation.

salient crack patterns, which are well aligned with ground truth in column (f). Column (j) shows the NMS based object detection results from our proposed approach, which are perceptually precise. It demonstrates the superior performance of our proposed *Crack-DA* in the UDA for the crack recognition.

Our proposed *Crack-DA* can be integrated seamlessly into various network backbones and boost the segmentation performance. As shown in Table I, we have also done experiments by integrating the proposed *Crack-DA* with various network backbones including Crack-Net [9], DeepCrack [13], DeepLabV3 [8], PSPNet [49], ASPP-Net [50], and Seg-Former [51] to test the performance of the proposed domain adaptation approach. The evaluation metrics follow the previous work [10] in pixel-level crack segmentation. It can be seen that our proposed *Crack-DA* boosts the segmentation performances on the target dataset by a large margin of about or more than 30% mean intersection over union (MIoU), which demonstrates its great effectiveness in increasing the domain generalization capacity. We directly do inference on the UAV onboard NVIDIA Xavier GPU to test the efficiency. We use the inference time of 6000×4000 images for MaWan-cracks and 512×512 for highway-cracks to evaluate the computational cost. It shows that for various network backbones, *Crack-DA* merely results in a marginal increase in the computational cost of less than 10%, which demonstrates that *Crack-DA* achieves great segmentation performance with satisfactory network efficiency. This can be explained by the fact that our proposed network optimization function is merely required in the training of the backbone network with *Crack-DA*. Once the training is finished, network weights are fixed in testing, and computations in proposed losses are not needed. Also, our proposed A-T is effective and light-weighted, only resulting in a marginal increment in computational costs for various network backbones.

Depth Estimation: We choose the off-the-shelf depth transformer [39] trained on KITTI [40], [41] to obtain depth pseudo labels. The training details on the KITTI dataset follow the original implementation [39]. As shown in Fig. 7, the depth transformer can offer great performance on the unseen test set in the autonomous driving scenarios with an accurate estimation of the geometric depth information. The depth estimation results by

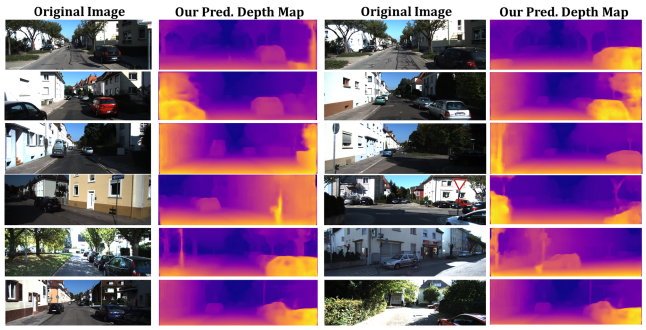

Fig. 7. Network prediction results of the depth estimation by trained models on the test set of KITTI Benchmark. Darker means deeper.

the source dataset depth decoder and target dataset decoder are shown in columns (b) and (c) of Fig. 6. It can be seen qualitatively that the depth estimated by the source decoder is less accurate than the target decoder and a large domain gap exists. Also, the cracks can be seen apparently and explicitly in column (c) for the target domain, proving that the model learns the crack patterns effectively in the depth estimation. For the results of the source domain in column (b), the crack patterns are clear although the depth is not accurate. It indicates that our designed A-T-based feature fusion network is important and the crack semantics and the depths are strongly coupled and correlated.

B. Experiments of UDA From Our Self-Established Dataset to the Highway Crack Dataset

We also did experiments of the UDA from our self-established dataset to the UAV-captured highway pavement crack dataset [53].

Experimental Settings: The experiment settings are the same as the test with MaWan crack. For the unlabeled set, we have used the 900 target-domain images with the resolutions of 512×512 in the UAV-captured highway pavement crack dataset [53] for the testing of domain adaptive segmentation.

Experimental Results: We have shown the qualitative crack segmentation experimental results in Fig. 8. The segmentation of

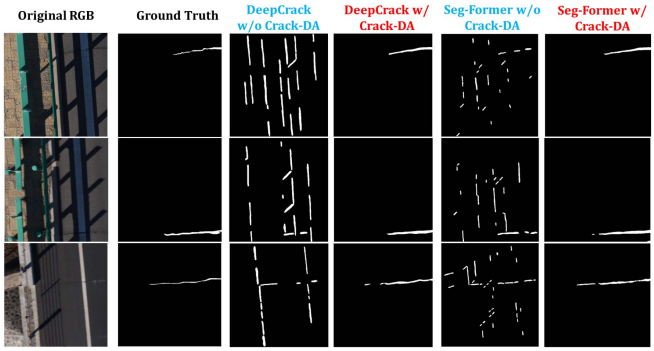

Fig. 8. Qualitative comparisons of w/ and w/o the proposed *Crack-DA* for the domain adaptive highway pavement crack segmentation.

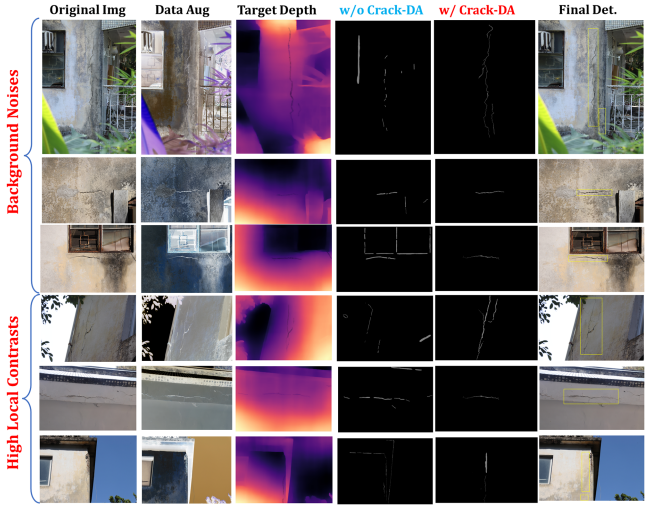

Fig. 9. Industrial inspection cases where the prevailing approach DeepCrack [13] will fail and our proposed *Crack-DA* can tackle the challenging circumstances well with excellent performance and robustness.

the UAV-captured pavement cracks is a tough task because of the dark lighting conditions and complex background objects with occlusions and shadows as shown in the first column of Fig. 8, causing a great domain gap in recognitions. The DeepCrack [13] and Seg-Former [51] are the previous fully supervised SOTAs method for crack segmentation and general segmentation. It can be seen that when the domain gap is large between the source and the target domain, previous SOTAs can not handle the complex backgrounds such as shadows and the road handrails. They deal poorly with the low-light conditions and give many false predictions when encountered with edges or various noises in the background, as shown in Fig. 8. The results with our proposed *Crack-DA* are shown in the fourth and sixth columns, respectively. It can be demonstrated that our proposed *Crack-DA* has a great boost on the segmentation performance. The intact crack patterns are successfully captured with explicit details. And the noncrack background is successfully suppressed. The cracks segmented become more distinctive with fine-grained details. It can be explained by the fact that our proposed edge and depth aware domain adaptive learning strategy can exploit the depth and edge information to suppress the background noncrack patterns and explicitly extract the detailed crack patterns. The quantitative results with and without *Crack-DA* are shown in Table I. With *Crack-DA*, the segmentation performance can be largely enhanced for various network architectures.

C. Discussions About Industrial Applications

1) Typical Challenging Cases: In real industrial applications of UAV-based infrastructural inspections of MaWan buildings, the current SOTAs learning-based approaches such as DeepCrack [13] will fail in two main circumstances. As shown in Fig. 9, the first circumstance is when faced with complex background noises or objects, such as the air conditioners, the grasses, the windows, and planks, etc. The second circumstance is when faced with high local contrasts in the local gradient of pixels. The current deep learning-based approaches including the DeepCrack [13] and SegFormer [51] have poor performance in such cases for the fact that these methods can not well capture the geometric depth features and suppress the noncrack background noises within images.

However, our approach *Crack-DA* can handle those problems well owe to our two proposed modules to improve the domain adaptation capacity of the network. The first module is our proposed *optimization functions with edge and depth awareness*. As shown from our experimental results in Figs. 6 and 9, cracks can be seen very apparently from both the edge and depth maps. The current crack recognition method have long overlooked the important edge and depth information in crack recognition. The edge shows explicitly, where the cracks can possibly emerge. And the geometric depth changes reveal cracks in 3-D structures. From principle, cracks can merely appear at the pixels that have a large gradient in the neighborhood. Our design edge-aware loss encourages the network to predict pixels with a small local gradient as noncracks to suppress the noise and noncrack patterns. According to our ablation studies in the next subsection, both the depth-aware loss L_{Dep} and the edge-aware loss L_{Edge} have significant boost in the crack recognition performance in the real-world industrial circumstances of MaWan building inspections and pavement inspections.

The second module is our proposed *domain gap estimation by depth*. By using the differences in self-supervised depth predictions to estimate the domain gap and formulate the domain shift into the network optimizations, the influence of domain shift to the segmentation performance can be significantly eased according to our experiments. The depth and semantic are inherently coupled and we have explicitly modeled their correlations by our proposed transformer-based feature fusion network. Therefore, the high value of discrepancy in depth estimation implies a great domain shift across domains. Accordingly, we should assign a small weight u in the domain adaptive semantic segmentation loss L_{Seg}. According to our ablation experimental results in the next section, our proposed *domain gap estimation by depth* module has a prominent boost on the real-world recognition performance.

Fig. 10. Influence of thresholds on the edge detection.

***2) Key Parameters of* Crack-DA:** According to our substantial experiments conducted in the above two sections, some key parameters should be set properly to achieve satisfactory performance in domain adaptations. In the following, we discuss the influences of key parameters in each module of our proposed UDA framework.

Edge Detection: First and foremost, weights in edge detectors should be set carefully because the quality of the edge map is very significant to suppress the background noises and the noncrack patterns. According to our experimental results shown in Fig. 9, if the Canny [42] edge detector is adopted, we should set the edge segmentation threshold τ in the appropriate range to guarantee the edge is explicitly extracted and the background noises are effectively eliminated. Also, as is explicitly and apparently shown in our experimental results in Fig. 10, a too large threshold in edge detection resulted in vague edge patterns, and a too small threshold resulted in severe background noises retained. Both the blurred edge patterns and the retained background noises will have detrimental effects on the final edge-guided optimization functions. Therefore, a modest threshold $\tau \in [0.4, 0.6]$ should be set to guarantee a high-quality edge map.

Segmentation Loss L_{Seg}: The use of focal loss implies we should assign larger weights to the minority crack pixels, and assign small weights to the majority of noncrack pixels. We set the hyperparameter α in the focal loss. According to our experiments, the robustness can be guaranteed if set α in the range 2–5. A too large α resulted in over-fitting of the network to the hard-to-classified minority class, which is the crack in our circumstance. And a too small α resulted in the convergence of the network in the early training stages, which often means the optimization was stuck into the local minimum. Note that the setting of α is very significant in the training network in the industrial applications, because the defects or anomalies in industrial applications are often the minority classes that are hard to be classified and the weights in the focal loss should be set in the proper range as indicated ($\alpha \in [2, 5]$).

Self-Supervised Depth Loss L_{dep}: For our proposed self-supervised depth loss L_{dep}, we have a hyperparameter a is the depth threshold set to $\frac{1}{5}$ of the maximum depth difference in the pseudoground truth depth map. A too large ($\frac{1}{2}$ of the maximum difference) or too small a ($\frac{1}{20}$ of the maximum difference) will result in inappropriate penalization in depth prediction, which will slightly influence the final segmentation performance. However, according to our tests for the MaWan building cracks, this will merely result in the performance drop in MIoU of less than 0.2%, and thus negligible. In other words, our proposed self-supervised depth loss L_{dep} is robust to the changing of hyperparameters within an acceptable range.

In summary, setting appropriate key parameters is very significant to the robust network performance in challenging real-world industrial circumstances. Therefore, we have included above-illustrated key parameters to facilitate the hyperparameter settings in future industrial applications for the research community.

D. Ablation Studies of the Proposed Crack-DA

1) Ablations:

a) Without specific optimization function terms or weights: We have also done ablation experiments for different network modules. We have ablated the network modules in all settings listed as follows: 1) Remove the L_{Aug} for data augmentation. 2) Removing L_{Edge}, which means that we remove the guidance from the edge detection. 3) Removing L_{Dep}, which means that we remove the guidance from the depth estimation. 4) Removing the domain adaptation weight u in L^T_{Seg} for the target dataset, which means that we do not consider the domain gap in the training of the target dataset. 5) Keeping all the network modules.

b) Without the A-T-based feature fusion network: We substitute the A-T-based feature fusion network in Fig. 5 with four MLPs without attention between the depth feature and semantic feature. Also, we replace the proposed A-T with the SOTAs swin-transformer [54] to test the performance.

2) Results: The results are shown in Table II. For MaWan building cracks, it can be seen that if dropping the data augmentation loss L_{Aug}, the MIoU will drop by 3.9%, which demonstrates the effectiveness of proposed data augmentation strategies to enrich and boost the training samples. Also, without the edge-aware loss term L_{Edge}, the segmentation MIoU on the target dataset will drop significantly by 5.3%. It demonstrates that edge information is of great significance to help the network precisely locate the nonedge regions and eliminate the background noises. In addition, when without the depth aware loss term L_{Dep}, the MIoU drops by a large margin of 6.9%, which demonstrates that learning the geometric depth information of visual objects, especially cracks in the image, is of great significance to inferring the semantics. Finally, when tested without the weight u in L^T_{Seg}, the segmentation MIoU drops most by 13.5%. It demonstrates that considering the domain gap in network training is important to boost the performance on the target test set. It also implies that the domain gap can be successfully estimated by the geometric depth estimation difference, which validates our hypothesis

TABLE II
ABLATION STUDY ON UDA TASKS OF HIGHWAY AND MAWAN BUILDING CRACKS

Ablation Study Setting	Ablation Target	Highway Pavement Cracks					MaWan Building Cracks				
		GA/%	**MIoU**/%	Precision/%	Recall/%	F-Score/%	GA/%	**MIoU**/%	Precision/%	Recall/%	F-Score/%
1) w/o L_{Aug}	Loss Term 1	82.5 (↓3.7)	72.6 (↓6.3)	72.8 (↓6.0)	73.1 (↓5.8)	72.9 (↓6.0)	77.6 (↓5.9)	72.3 (↓3.9)	72.2 (↓2.1)	72.9 (↓2.3)	72.5 (↓2.2)
1) w/o L_{Edge}	Loss Term 2	81.5 (↓4.7)	71.2 (↓7.7)	72.5 (↓6.3)	74.5 (↓4.4)	73.5 (↓5.4)	75.5 (↓8.0)	70.9 (↓5.3)	70.9 (↓3.4)	70.5 (↓4.7)	70.7 (↓4.0)
1) w/o L_{Dep}	Loss Term 3	80.3 (↓5.9)	71.8 (↓7.1)	72.3 (↓6.5)	72.8 (↓6.1)	72.6 (↓6.3)	76.2 (↓7.3)	69.3 (↓6.9)	69.9 (↓4.4)	68.5 (↓6.7)	69.2 (↓5.5)
1) w/o the weight u in L^T_{Seg}	Loss Term Weights	72.1 (↓14.1)	63.2 (↓15.7)	62.9 (↓15.9)	64.2 (↓14.7)	63.5 (↓15.4)	69.7 (↓13.8)	62.7 (↓13.5)	62.9 (↓11.4)	62.3 (↓12.9)	62.6 (↓12.1)
2) Replace A-T with MLPs	Network Structure	79.5 (↓6.7)	71.3 (↓7.6)	72.6 (↓6.2)	68.5 (↓10.4)	70.6 (↓8.3)	77.8 (↓5.7)	69.1 (↓7.1)	66.3 (↓8.0)	71.5 (↓3.7)	68.8 (↓5.9)
2) Replace A-T with Swin-Transformer	Network Structure	86.2 (↑0.0)	78.8 (↓0.1)	79.1 (↑0.3)	78.7 (↓0.2)	78.9 (↑0.0)	83.4 (↓0.1)	76.5 (↑0.3)	74.3 (↑0.0)	75.5 (↑0.3)	74.9 (↑0.2)
3) **Baseline Seg-Former w/ *Crack-DA***	No Ablation	**86.2**	**78.9**	**78.8**	**78.9**	**78.9**	**83.5**	**76.2**	**74.3**	**75.2**	**74.7**

We test Crack-DA with the best backbone Seg-Former.

that semantics and geometric depth are closely related. For the ablations in the network structure, we can see that the attentional transformer (A-T) is also important to network performance. Not considering the relationship between depth and semantics will result in a large MIoU drop of 7.1%. However, when replacing the attentional transformer proposed with the current SOTAs Swin-Transformer [54], the network performance will maintain at the same level (↑0.3%). It demonstrates that a simple network structure of matrices with weight is enough in our case to model the relations between depth and semantics. It to some extent implies that the relationship between depth and semantics is easy to be learned with a proposed simple A-T. In summary, it is demonstrated that all the proposed optimization function terms and network components in the proposed *Crack-DA* are significant to the UDA performance.

V. CONCLUSION

In this work, we have proposed a systematic framework to solve UAV-based industrial unsupervised domain adaptive crack inspections. First, we have systematically designed the autonomous UAV to conduct the inspection task and developed related 3-D reconstruction algorithms for the target infrastructure. We have then developed a *Crack-DA* framework for the domain adaptive crack segmentation and the subsequent detection, which achieves depth and edge awareness with our proposed fusion network structures and optimization functions. Extensive experimental results demonstrate that our proposed *Crack-DA* performance in domain adaptation with labels only from our self-established crack segmentation database. To the best of our knowledge, our proposed method is the first attempt that achieves satisfactory performance in domain adaptive real-site industrial crack recognitions without any manual labeling for the inspected target. In conclusion, we have proposed the *Crack-DA* for domain adaptive crack recognitions with great effectiveness and robustness. Our proposed frameworks are of great significance to the development of the smart city with autonomous systems to perform domain adaptive inspections of infrastructural cracks.

REFERENCES

[1] W. Choi and Y.-J. Cha, "SDDNet: Real-time crack segmentation," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8016–8025, Sep. 2020.

[2] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and naïve bayes data fusion," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4392–4400, May 2018.

[3] L. Wang and Z. Zhang, "Automatic detection of wind turbine blade surface cracks based on UAV-taken images," *IEEE Trans. Ind. Electron.*, vol. 64, no. 9, pp. 7293–7303, Sep. 2017.

[4] L. Wen, X. Li, and L. Gao, "A new reinforcement learning based learning rate scheduler for convolutional neural network in fault classification," *IEEE Trans. Ind. Electron.*, vol. 68, no. 12, pp. 12890–12900, Dec. 2021.

[5] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[6] N. Wang, Z. Zhang, X. Zhao, Q. Miao, R. Ji, and Y. Gao, "Exploring high-order correlations for industry anomaly detection," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9682–9691, Dec. 2019.

[7] J. Yang, C. Zhou, S. Yang, H. Xu, and B. Hu, "Anomaly detection based on zone partition for security protection of industrial cyber-physical systems," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4257–4267, May 2018.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[9] K. Liu, X. Han, and B. M. Chen, "Deep learning based automatic crack detection and segmentation for unmanned aerial vehicle inspections," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2019, pp. 381–387.

[10] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "Deepcrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019.

[11] K. Liu, Z. Gao, F. Lin, and B. M. Chen, "FG-Conv: Large-scale LiDAR point clouds understanding leveraging feature correlation mining and geometric-aware modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 12896–12902.

[12] Z. Liu, X. Liu, Z. Cao, X. Gong, M. Tan, and J. Yu, "High precision calibration for 3D vision-guided robot system," *IEEE Trans. Ind. Electron.*, vol. 70, no. 1, pp. 624–634, Jan. 2023.

[13] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "DeepCrack: Learning hierarchical convolutional features for crack detection," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1498–1512, Mar. 2019.

[14] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "Cracktree : Automatic crack detection from pavement images," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 227–238, 2012.

[15] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 9592–9600.

[16] G. Zhang, K. Cui, T.-Y. Hung, and S. Lu, "Defect-GAN: High-fidelity defect synthesis for automated defect inspection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2524–2534.

[17] M. Mundt, S. Majumder, S. Murali, P. Panetsos, and V. Ramesh, "Meta-learning convolutional neural architectures for multi-target concrete defect classification with the concrete defect bridge image dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11196–11 205.

[18] Q. Wan, et al., "Industrial image anomaly localization based on gaussian clustering of pre-trained feature," *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6182–6192, Jun. 2022.

[19] P. Sassi, P. Tripicchio, and C. A. Avizzano, "A smart monitoring system for automatic welding defect detection," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9641–9650, Dec. 2019.

[20] K. Liu, Y. Zhao, Z. Gao, and B. M. Chen, "WeakLabel3D-Net: A complete framework for real-scene LiDAR point clouds weakly supervised multi-tasks understanding," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 5108–5115.

[21] K. Liu, Y. Zhao, Q. Nie, Z. Gao, and B. M. Chen, "Weakly supervised 3D scene segmentation with region-level boundary awareness and instance discrimination," in *Eur. Conf. Comput. Vis.*, 2022.

[22] J. Huang, D. Guan, A. Xiao, and S. Lu, "Multi-level adversarial network for domain adaptive semantic segmentation," *Pattern Recognit.*, vol. 123, 2022, Art. no. 108384.

[23] J. Huang, D. Guan, A. Xiao, and S. Lu, "Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 3635–3649, 2021.
[24] F. Zhan, C. Xue, and S. Lu, "GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9105–9115.
[25] C. Zhang et al., "Self-guided adaptation: Progressive representation alignment for domain adaptive object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 2246–2258, May 2021.
[26] J. Huang, S. Lu, D. Guan, and X. Zhang, "Contextual-relation consistent domain adaptation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 705–722.
[27] J. Huang, D. Guan, A. Xiao, and S. Lu, "RDA: Robust domain adaptation via fourier adversarial attacking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8988–8999.
[28] Z. Luo et al., "Unsupervised domain adaptive 3D detection with multi-level consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8866–8875.
[29] D. Guan, J. Huang, A. Xiao, and S. Lu, "Domain adaptive video segmentation via temporal consistency regularization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8053–8064.
[30] J. Huang, D. Guan, A. Xiao, S. Lu, and L. Shao, "Category contrast for unsupervised domain adaptation in visual tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1203–1214.
[31] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8827–8836.
[32] M. Zhen et al., "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13666–13675.
[33] Z. Li et al., "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6197–6206.
[34] M.-J. Yang, Y.-X. Guo, B. Zhou, and X. Tong, "Indoor scene generation from a collection of semantic-segmented depth images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15203–15212.
[35] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4545–4554.
[36] M. Mandal, L. K. Kumar, and S. K. Vipparthi, "MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2626–2635.
[37] C. YuanQiang et al., "Guided attention network for object detection and counting on drones," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 709–717.
[38] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *Proc. Int. Workshop Reproducible Res. Pattern Recognit.*, Springer, 2016, pp. 60–74.
[39] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16269–16279.
[40] J. Behley et al., "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9297–9307.
[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[42] P. Bao, L. Zhang, and X. Wu, "Canny edge detection enhancement by scale multiplication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1485–1490, Sep. 2005.
[43] Q. Wen et al., "Time series data augmentation for deep learning: A survey," 2020, *arXiv:2002.12478*.
[44] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[45] S. Lambert-Lacroix and L. Zwald, "The adaptive Berhu penalty in robust regression," *J. Nonparametric Statist.*, vol. 28, no. 3, pp. 487–514, 2016.
[46] K. Liu, X. Zhou, and B. M. Chen, "An enhanced lidar inertial localization and mapping system for unmanned ground vehicles," in *Proc. IEEE 17th Int. Conf. Control Automat.*, 2022, pp. 587–592.
[47] K. Liu, X. Zhou, B. Zhao, H. Ou, and B. M. Chen, "An integrated visual system for unmanned aerial vehicles following ground vehicles: Simulations and experiments," in *Proc. IEEE 17th Int. Conf. Control Automat.*, 2022, pp. 593–598.
[48] Z. Liu, X. Qi, and C.-W. Fu, "One thing one click: A self-training approach for weakly supervised 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1726–1736.
[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 2881–2890.
[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[51] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.
[52] Z. Zhang, S. Gao, and Z. Huang, "An automatic glioma segmentation system using a multilevel attention pyramid scene parsing network," *Curr. Med. Imag.*, vol. 17, no. 6, pp. 751–761, 2021.
[53] Z. Hong et al., "Highway crack segmentation from unmanned aerial vehicle images using deep learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art no. 6503405.
[54] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

Kangcheng Liu (Member, IEEE) received the B.Eng. degree in electrical and automation engineering from Harbin Institute of Technology, Harbin, China, in 2018, and the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2022.

His research fields include robotics, LIDAR-SLAM, 3D Vision, and robot control.

Dr. Liu serves as a Reviewers and has publications in IEEE International Conference on Robotics and Automation and several IEEE Transactions.

Ben M. Chen (Fellow, IEEE) is currently a Professor of mechanical and automation engineering with the Chinese University of Hong Kong (CUHK), Hong Kong. He was a Provost's Chair Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), before joining CUHK, in 2018. He was an Assistant Professor with the Department of Electrical Engineering, State University of New York at Stony Brook, NY, USA, from 1992 to 1993. He has authored/co-authored hundreds of journal and conference articles, and a dozen research monographs in control theory and applications, unmanned systems and financial market modeling. His current research interests are in unmanned systems and their applications.

Dr. Chen is a Fellow of IEEE and Fellow of Academy of Engineering, Singapore. He had served on the editorial boards of a dozen international journals including Automatica and IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He is currently serving as an Editor-in-Chief of Unmanned Systems, and Editor of *International Journal of Robust and Nonlinear Control*.