Mind marginal non-crack regions: Clustering-inspired representation learning for crack segmentation

Zhuangzhuang Chen, Zhuonan Lai, Jie Chen, Jianqiang Li *
Shenzhen University, Shenzhen, China
{chenzhuangzhuang2016, 2018101068}@email.szu.edu.cn
{chenjie, lijq}@szu.edu.cn

Abstract

Crack segmentation datasets make great efforts to obtain the ground truth crack or non-crack labels as clearly as possible. However, it can be observed that ambiguities are still inevitable when considering the marginal non-crack region, due to low contrast and heterogeneous texture. To solve this problem, we propose a novel clustering-inspired representation learning framework, which contains a two-phase strategy for automatic crack segmentation. In the first phase, a pre-process is proposed to localize the marginal non-crack region. Then, we propose an ambiguity-aware segmentation loss (Aseg Loss) that enables crack segmentation models to capture ambiguities in the above regions via learning segmentation variance, which allows us to further localize ambiguous regions. In the second phase, to learn the discriminative features of the above regions, we propose a clustering-inspired loss (CI Loss) that alters the supervision learning of these regions into an unsupervised clustering manner. We demonstrate that the proposed method could surpass the existing crack segmentation models on various datasets and our constructed CrackSeg5k dataset.

1. Introduction

Concrete structure health monitoring plays an essential role in industrial scenarios [24, 29, 32, 45], of which crack segmentation is the last and indispensable stage. With aging concrete structures, the need for maintenance increases, which would lead to poor health conditions or structural deficiencies if not addressed properly [20, 22]. For this reason, we argue that it is necessary to repair cracks before the onset of serious deterioration so as to lighten the manual maintenance burden [5]. However, the ambiguity in marginal non-crack regions is still inevitable due to low contrast, heterogeneous texture, and the uncertainty of the segmented region without domain knowledge. Considering this, pixel-level segmentation of cracks remains challenging [3, 7, 63].

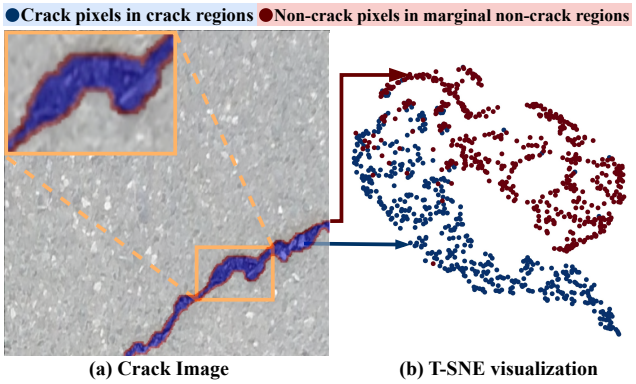

Figure 1. The ambiguity problem in marginal non-crack regions. (a) Blue indicates crack regions and red is marginal non-crack regions. Due to the unclear crack boundary, the labels of marginal non-crack regions are likely to be ambiguous (a close look at the marginal region is present in the orange box). (b) T-SNE feature visualizations of the crack pixels (blue) in crack regions and non-crack pixels (red) in marginal non-crack regions.

Nowadays, substantial progress of deep learning techniques also promotes research on crack segmentation tasks [8, 32, 41]. For instance, DeepCrack [67] proposes a multi-stage fusion for crack segmentation, which is derived from the commonly used encoder-decoder architecture (SegNet [2]). Inspired by ViT [51], Crackformer [32] proposes a crack transformer network to capture long-range interactions for fine-grained crack segmentation. By exploiting boundary information, JTFN [8] exploits crack boundary as an additional supervision for crack segmentation tasks. However, theses methods have a crucial limitation: *the inability to extract discriminative features due to ambiguity in marginal non-crack regions.* That is, unlike objects in the natural image, there might be no salient structure boundary due to low contrast between crack and non-crack regions. Thus, it is challenging for crack segmentation mod-

*Corresponding Author Email: lijq@szu.edu.cn

els to extract the discriminative features, when encountering ambiguous labels in marginal non-crack regions. To verify this, Fig. 1 (b) visualizes the features of image pixels within crack regions and marginal non-crack regions. It can be observed that each class of pixels projected in the feature space spreads over a wide area, and crack/non-crack features are mixed and close to the class boundary. Therefore, unlike the previous methods that only focus on learning crack boundaries, we argue that there is room for improving the crack segmentation performance by addressing ambiguity in marginal non-crack regions.

In this paper, we propose a clustering-inspired representation learning framework, called CIRL, to address the above problem in a two-phase manner. Specifically, in the first phase, we first localize marginal non-crack regions in our pre-process. Then, to capture ambiguity in the above regions, we propose a novel ambiguity-aware segmentation loss, namely Aseg Loss, which is based on Wasserstein distances [1] for learning segmentation variance. With the help of learned variance, we are allowed to further determine ambiguous regions. Intuitively, these ambiguous regions are likely to have ambiguous labels, which makes it hard for the crack segmentation model to learn the discriminative feature. Considering this, in the second phase, we start a new perspective that alters the existing supervised learning for ambiguous regions as an unsupervised clustering manner, and draw inspiration from the consensus: *local neighbors in the feature space are more likely involved in the same cluster and should have more similar predictions than other features [10, 59].* Based on these, we propose a clustering-inspired loss, namely CI Loss, which aims to let local neighbor features involve similar predictions while features farther away have dissimilar predictions. Notably, the proposed method can be conveniently implemented on the top of most crack segmentation models. Moreover, the flexibility of the proposed method can promote crack segmentation models to more precisely segment crack regions by integrating Aseg and CI Loss. In summary, our main contributions include:

- To the best of our knowledge, this is the first work to consider ambiguity in marginal non-crack regions. For this cause, this paper proposes a two-phase clustering-inspired representation learning (CIRL) framework to escape from the disturbance of those ambiguous labels.
- Propose an ambiguity-aware segmentation loss (Aseg Loss) that enables crack segmentation models to capture ambiguity in marginal non-crack regions via learning segmentation variances.
- A clustering-inspired loss (CI Loss) achieves the goal of CIRL by altering the existing supervised learning into ambiguous regions as an unsupervised clustering manner.
- Extensive experiments verify the superiority of the proposed methods on public datasets and our constructed CrackSeg5k dataset.

2. Related work

2.1. Crack segmentation

With the great progress of deep models on nature image segmentation tasks, researchers make great efforts to apply these models to crack segmentation tasks [29, 36, 41, 58, 65]. For example, considering the great success of fully convolutional networks (FCN) [36], the variant of FCN has a surge in achieving end-to-end network training for crack segmentation tasks. Typically, Yang et al. [60] exploit the VGG-19 network[49] as the decoder, and integrate the FCN model for crack segmentation in pavements and concrete walls. Subsequently, Wang et al. [54] design a new FCN-based model for obtaining local information in crack segmentation tasks. Moreover, DeepCrack [67] shows its superiority over the existing methods, which adopt an encoder-decoder network by following the SegNet [2] for crack segmentation in various real-world scenarios. Later, U-Net [46] attracts widespread attention in the development of crack segmentation tasks due to its simplicity. Specifically, Liu et al. [35] prove that U-Net is more elegant, robust, and effective than other deep models for crack segmentation tasks. By combining with squeeze and excitation module [18], existing work [33] implements an improved U-Net with a pre-trained ResNet-34 [16]. Since ViT [51] shows its great power in general semantic segmentation tasks, Crackformer [32] is proposed to integrate self-attention mechanism for capturing long-range interactions for fine-grained crack segmentation tasks. Moreover, JTFN [8] and BACS [31] exploit crack edge as an additional supervision for crack segmentation. In addition, Marmanis et al. [38] and Liu et al. [34] integrate semantic segmentation and edge detection to further reduce the semantic ambiguity in remote sensing tasks. Similarly, Zhou et al. [64] employ boundary information as an additional assistance for producing more consistent outputs.

2.2. Crack segmentation loss

To better learn the discriminative features, different kinds of losses are proposed for crack segmentation. DeepCrack [67] designs the multi-scale cross-entropy losses to enhance the model to extract discriminative features at different scales. Moreover, previous studies [12, 43] reveal that dice Loss and focal Loss have advantages in crack segmentation tasks. Both of them can help the crack segmentation model to overcome the imbalance problem between crack and non-crack pixels and learn a better feature space, so that different pixels can be successfully classified. Further, a pixel-based adaptive weighted cross-entropy loss in conjunction with Jaccard distance is proposed to facilitate high-quality pixel-level road crack segmentation applications [26]. To

preserve the continuity of cracks, Pantoja-Rosero et al. [44] design a new connectivity-oriented loss by considering a more reasonable crack topology.

2.3. Deep clustering learning

Due to the fact that our method aims to perform unsupervised clustering learning for ambiguous regions, we give a brief review of related research on this scope. Recent deep clustering methods can be roughly divided into two groups: alternately or simultaneously learning the feature representation and cluster assignments. In the first group, DAC [4] and DCCM [56] can serve as typical examples that alternately update cluster assignments and between-sample similarity. Many methods in the second group try to maximize mutual information between samples and their augmentations [10]. Inspired by contrastive learning, many unsupervised clustering works [9, 62] combine InfoNCE [42] to learn a better feature space. Interestingly, NNCLR [13] provides a new manner that uses nearest neighbors in the latent space as positives in contrastive learning. However, it suffers from those negative pairs that may contain the samples from the same class. Besides, since it performs augmentation at the image level, it is hard to directly be applied to pixel-level crack segmentation tasks.

Our proposed CIRL shares the same purpose as the previous methods that focus on ambiguity in crack segmentation. However, ours is distinct from these methods in threefold: (1) This paper focuses on marginal non-crack regions, instead of crack boundary. Thus, our method is less sensitive than edge-based or boundary key point-based methods when facing annotation errors. (2) The proposed Aseg Loss can help our network learn segmentation variances of marginal non-crack regions, which can further localize ambiguous regions. (3) The proposed CI Loss addresses ambiguous region problems in pixel-level crack segmentation tasks, and it constructs two features sets for unsupervised clustering learning, instead of positive and negative pairs in existing contrastive clustering learning.

3. Method

In this section, we introduce the proposed two-phase CIRL that contains two consecutive phases. Specifically, we start with the motivation of CIRL in Sec. 3.1. Next, the first phase focuses on learning the ambiguity in marginal non-crack regions, so as to localize the ambiguous regions (Sec. 3.2). The second phase then alters supervised learning into unsupervised clustering learning for ambiguous regions (Sec. 3.3). Fig. 3 provides an overview of our approach.

3.1. Motivation

CIRL starts from an intuitive idea: due to the low contrast between crack and non-crack regions, there is ambiguity in marginal non-crack regions. Thus, the labels of these regions are likely to be unclear, making it hard for existing crack segmentation models to learn discriminative features.

Sub-training sets	1	2	3	4
Performance gap (F1-score)	-3.12 %	3.48 %	2.93%	1.79%

Table 1. The performance gaps come from the training of including/excluding marginal non-crack regions on four sub-training sets randomly splited from the training set of CrackSeg5k. The above results are obtained from the same test set in CrackSeg5k.

Image GT KL Our

Figure 2. The learned variance map by KL Divergence and CIRL.

To verify this intuition, we conduct an experiment on the CrackSeg5k dataset. Specifically, we first train Crackformer [32] on CrackSeg5k under the supervision of Binary Cross Entropy (BCE) Loss [11], and extract the features of image pixels within crack regions and marginal non-crack regions from the randomly selected training samples. In our pre-process, given the ground-truth crack segmentation map y^{gt}, the marginal non-crack region map M is obtained via dilation operations in Opencv[1], shown as follows:

$$M = dilate(y^{gt}) - y^{gt}, \tag{1}$$

where $dilate(\cdot)$ denotes dilation operations with the kernel size of 5×5. After that, we use t-SNE [50] to visualize the features of crack and non-crack pixels with different colors in Fig. 1 (b). It can be observed that some crack pixels and non-crack pixels are entangled in the feature space. The reason behind this is that some pixels of marginal non-crack regions have the same appearance as crack pixels. Thus, it is unwise to directly exploit the labels of these regions when training crack segmentation models.

With the above discussion, we now provide an additional experiment to reveal that a well-designed solution is needed for marginal non-crack regions in crack segmentation tasks. Firstly, we randomly split the training set of CrackSeg5k into four sub-training sets. Then, we exploit the above sub-training sets to train Crackformer with the supervision of BCE Loss in two manners: 1) exploiting the labels of marginal non-crack regions for training. 2) excluding marginal non-crack regions from training by referring to Eq. 1. Table 2 shows the inconsistent performance gap between two manners. The reason behind this is that some ambiguous regions may have a negative effect by directly

[1]https://opencv.org/

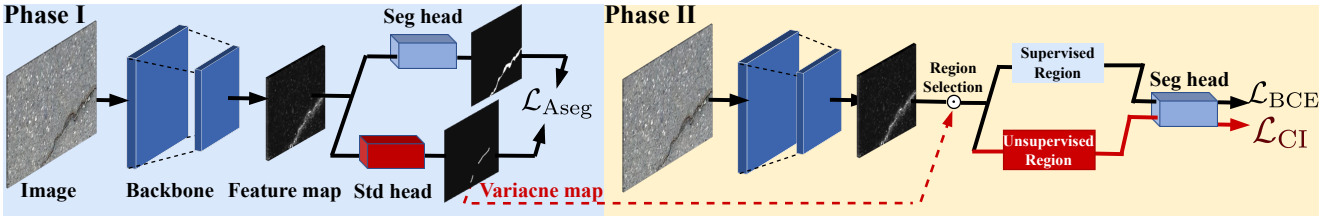

Figure 3. Our method contains two sequential phases: (1) Our network first introduces a std head for estimating standard deviations along with segmentation under the supervision of our ambiguity-aware segmentation loss ($\mathcal{L}_{\text{Aseg}}$). Then, the learned variance map help us to localize ambiguous regions in marginal non-crack regions. (2) We further alter the existing supervised learning problem for these ambiguous regions into an unsupervised clustering manner. Then, a clustering-inspired loss ($\mathcal{L}_{\text{CI}}$) is proposed to help our network escape from the disturbance of those ambiguous labels and learn discriminative features. The Std head is removed, and only the backbone and Seg head are needed for prediction. Thus, our framework does not introduce extra computation and memory at inference.

using their labels in the training of crack segmentation models, while the other regions can improve the above models by providing more training samples at the pixel level. To better understand this, the visualization of our localized ambiguous regions can be found in Fig 2. *Intuitively, ambiguous regions in marginal non-crack regions should be further localized and treated separately from the other regions.*

3.2. Phase I: Learning the ambiguity in marginal non-crack regions

Based on the previous discussion, in the first stage, we aim to estimate the segmentation confidence along with segmentation, so as to capture ambiguity in marginal non-crack regions. To achieve this, for the pixel at the position (i, j) of the input image, our network predicts a probability distribution $P^{\Theta}_{i,j}(y)$ instead of one single label. Herein, we simply assume that each predicted label (i.e., single-variate) at the pixel level obeys an independent Gaussian distribution. And then, we have the following equation:

$$P^{\Theta}_{i,j}(y) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{i,j}}} e^{-\frac{\left(y-y^p_{i,j}\right)^2}{2\hat{\sigma}^2_{i,j}}} \tag{2}$$

where Θ is the set of learnable parameters including backbone, seg head, and std head, shown in Fig. 3. $y^p_{i,j}$ and $\hat{\sigma}_{i,j}$ denotes the predicted label and estimated standard deviation at the position (i, j) of an input image, respectively. When $\hat{\sigma}_{i,j}$ is very close to 0, it indicates that our network has a large confidence in the current predicted label. Note that, our std head is implemented by a fully-connected layer on top of the backbone, as shown in Fig. 3. Accordingly, the corresponding ground-truth label $y^{gt}_{i,j}$ can also be formulated as a Gaussian distribution $\mathcal{N}(y^{gt}_{i,j}, \sigma^2_{i,j})$ with its standard deviation $\sigma_{i,j} \to 0$. Then, this gaussian distribution can be viewed as: $P^{gt}_{i,j}(y) = \delta\left(y - y^{gt}_{i,j}\right)$, where $\delta(\cdot)$ indicates Dirac delta function. And then, the proposed Aseg Loss can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{Aseg}} &= \sum_{i=1}^{H}\sum_{j=1}^{W} \frac{D_W\left(P^{\Theta}_{i,j}(y) \| P^{gt}_{i,j}(y)\right)}{\lambda + \hat{\sigma}^2_{i,j}} \\ &= \sum_{i=1}^{H}\sum_{j=1}^{W} \frac{\left\|y^p_{i,j} - y^{gt}_{i,j}\right\|_2^2 + \hat{\sigma}^2_{i,j}}{\lambda + \hat{\sigma}^2_{i,j}}\end{aligned} \tag{3}$$

where H and W indicate the height and width of an input image. The effect of the hyper-parameter λ will be explained later. Herein, we exploit Wasserstein distance as the distance metric $D_W(\cdot)$ and minimize the distance between $P^{\Theta}_{i,j}(y)$ and $P^{gt}_{i,j}(y)$. Meanwhile, $D_W\left(P^{\Theta}_{i,j}(y) \| P^{gt}_{i,j}(y)\right)$ can be unfolded by the following proposition:

Proposition 1 *As previously discussed, given that $P^{gt}_{i,j}(y)$ is the Dirac delta function as:*

$$\delta\left(y - y^{gt}_{i,j}\right) = \lim_{\mu \to y^{gt}_{i,j}, \Sigma \to 0} \mathcal{N}(\mu, \Sigma), \tag{4}$$

an easy-to-compute term for $D_W\left(P^{\Theta}_{i,j}(y) \| P^{gt}_{i,j}(y)\right)$ can be derived as the following equation:

$$D_W\left(P^{\Theta}_{i,j}(y) \| P^{gt}_{i,j}(y)\right) = \left\|y^p_{i,j} - y^{gt}_{i,j}\right\|_2^2 + \hat{\sigma}^2_{i,j}. \tag{5}$$

Proof 1 *Suppose that we have two multivariate Gaussians $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ in $\mathcal{R}^n$, then the Wasserstein distance between two distributions can be derived as follows:*

$$\begin{aligned}&W_2^2\left(\mathcal{N}\left(\mu_1, \Sigma_1\right), \mathcal{N}\left(\mu_2, \Sigma_2\right)\right) \\ &= \left\|\mu_1 - \mu_2\right\|^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\sqrt{\Sigma_1}\Sigma_2\sqrt{\Sigma_1}\right)^{\frac{1}{2}}\right)\end{aligned} \tag{6}$$

Now, by letting $\delta\left(y - y^{gt}_{i,j}\right)$ and $\mathcal{N}(y^p_{i,j}, \hat{\sigma}^2_{i,j})$ substitute in the above equation, we are allowed to obtain Proposition 1.

Notably, the reason for adopting the Wasserstein distance instead of KL-Divergence [17] is that the latter heavily relies on an indispensable intersection between two distributions [1]. In addition, when $y^p_{i,j}$ is predicted accurately,

i.e., $\left\|y_{i,j}^{p}-y_{i,j}^{gt}\right\| \rightarrow 0$, a smaller variance is expected to be produced by our our network. Based on this point, a term $\lambda+\hat{\sigma}_{i,j}^{2}$ is added in Eq. 3 for the following reason: due to the ambiguity in marginal non-crack regions, when the predicted label is not consistent with the ground-truth label, i.e., $\left\|y_{i,j}^{p}-y_{i,j}^{gt}\right\|_{2}^{2}>\lambda$, the minimization of $\mathcal{L}_{\text{Aseg}}$ will enforce our network to produce the larger variance $\hat{\sigma}_{i,j}^{2}$. For numerical stability, we actually predict the log variance $\hat{s}=\log\hat{\sigma}_{i,j}^{2}$ and reformulate Eq. 3 as follows:

$$\begin{aligned}\mathcal{L}_{\text{Aseg}} &= \sum_{i=1}^{H}\sum_{j=1}^{W}\frac{D_{W}\left(P_{i,j}^{\Theta}(y)\|P_{i,j}^{gt}(y)\right)}{\lambda+\exp(\hat{s})} \\ &= \sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\left\|y_{i,j}^{p}-y_{i,j}^{gt}\right\|_{2}^{2}+\exp(\hat{s})}{\lambda+\exp(\hat{s})}.\end{aligned} \tag{7}$$

Now, with the learned variance $\hat{\sigma}_{i,j}^{2}=\exp(\hat{s})$, an input pixel at the position (i,j) is grouped into ambiguous regions by the following equation:

$$A_{i,j}=\begin{cases}1, & \text{case } \hat{\sigma}_{i,j}^{2}>\gamma \text{ and } M_{i,j}==1\\ 0, & \text{case } \hat{\sigma}_{i,j}^{2}\leq\gamma \text{ or } M_{i,j}==0\end{cases}, \tag{8}$$

where γ is a hyper-parameter that serves as a threshold. According to Eq. 1, $M_{i,j}=1$ indicates that the current pixel belongs to marginal non-crack regions.

3.3. Phase II: From supervised learning to unsupervised learning for ambiguous regions

With the help of the previous stage, we are allowed to localize ambiguous regions, which are supposed to have ambiguous labels. Thus, an uncontrollable effect will be caused when directly using these labels in the training process. For this purpose, in the second stage, we proposed clustering-inspired loss (CI Loss), which works in an unsupervised features clustering manner with an intuitive idea: *features that are located close/farther in feature space should have consistent/inconsistent predictions.*

Given an input image $\mathcal{I}$, let $S_{\mathcal{I}}$ denote the feature set of all pixels that belong to the ambiguous region by following Eq. 8. $\mathcal{F}_m$ and $\mathcal{F}_n$ denote the features of two pixels in $S_{\mathcal{I}}$, while having corresponding predict probabilities p_m and p_n. Inspired by [14, 59], we define $p_{m,n}$ as the probability that $\mathcal{F}_m$ have a consistent prediction to $\mathcal{F}_n$:

$$p_{m,n}=\frac{e^{p_m^T p_n}}{\sum_{\mathcal{F}_q\in S_{\mathcal{I}}}e^{p_m^T p_q}}. \tag{9}$$

For each feature $\mathcal{F}_m$ in $S_{\mathcal{I}}$, we define two sets: close neighbor set $\mathcal{C}_m$ and farther feature set $\mathcal{O}_m$. The former selects K-nearest neighbors of $\mathcal{F}_m$ from $S_{\mathcal{I}}$ with cosine similarity as the distance metric. The latter is constructed by excluding $\mathcal{C}_m$ and $\mathcal{F}_m$ from $S_{\mathcal{I}}$. Returning to our motivation, for each feature $\mathcal{F}_m$, the features in $\mathcal{O}_m$ should have more notable inconsistent predictions than those in $\mathcal{C}_m$. Based on this, we define the likelihood function between $\mathcal{F}_m$ and $\mathcal{C}_m$:

$$P\left(\mathcal{C}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)=\prod_{\mathcal{F}_n\in\mathcal{C}_m}p_{m,n}=\prod_{\mathcal{F}_n\in\mathcal{C}_m}\frac{e^{p_m^T p_n}}{\sum_{\mathcal{F}_q\in S_{\mathcal{I}}}e^{p_m^T p_q}}, \tag{10}$$

where $\theta_{\mathcal{B}}$ and $\theta_{\mathcal{S}}$ denotes the parameters of the backbone and seg head in our network. Similarly, the likelihood function between $\mathcal{F}_m$ and $\mathcal{O}_m$ can be defined as follows:

$$P\left(\mathcal{O}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)=\prod_{\mathcal{F}_n\in\mathcal{O}_m}p_{m,n}=\prod_{\mathcal{F}_n\in\mathcal{O}_m}\frac{e^{p_m^T p_n}}{\sum_{\mathcal{F}_q\in S_{\mathcal{I}}}e^{p_m^T p_q}}. \tag{11}$$

Now, the goal of our clustering-inspired loss can be achieved by minimizing the following negative log-likelihood function:

$$\psi(\mathcal{C}_m,\mathcal{O}_m)=-log\frac{P\left(\mathcal{C}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)}{P\left(\mathcal{O}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)}. \tag{12}$$

Noting that, when $S_{\mathcal{I}}$ is very large, it is inefficient and even impractical to compute the above equation. Considering this, we derive an upper bound as an alternative by the following proposition.

Proposition 2 *Suppose that $|\mathcal{O}_m|$ is significantly larger than $|\mathcal{C}_m|$, then we have an upper bound of $\psi(\mathcal{C}_m,\mathcal{O}_m)$, given by*

$$\begin{aligned}&\psi(\mathcal{C}_m,\mathcal{O}_m)=-log\frac{P\left(\mathcal{C}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)}{P\left(\mathcal{O}_m \mid \mathcal{F}_m,\theta_{\mathcal{B}},\theta_{\mathcal{S}}\right)}\\ &\leq -\sum_{\mathcal{F}_n\in\mathcal{C}_m}p_m^T p_n+\frac{|\mathcal{C}_m|}{|\mathcal{O}_m|}\sum_{\mathcal{F}_k\in\mathcal{O}_m}p_m^T p_k+\left(|\mathcal{C}_m|-|\mathcal{O}_m|\right)\log|S_{\mathcal{I}}|\\ &=\overline{\psi}(\mathcal{C}_m,\mathcal{O}_m)\end{aligned} \tag{13}$$

Proof 2 *According to Eq. 10-12, we have $\psi(\mathcal{C}_m,\mathcal{O}_m)$:*

$$\begin{aligned}&=-\sum_{\mathcal{F}_n\in\mathcal{C}_m}p_m^T p_n+\sum_{\mathcal{F}_k\in\mathcal{O}_m}p_m^T p_k+\left(|\mathcal{C}_m|-|\mathcal{O}_m|\right)\log\left(\sum_{\mathcal{F}_q\in S_{\mathcal{I}}}e^{p_m^T p_q}\right)\\ &\leq -\sum_{\mathcal{F}_n\in\mathcal{C}_m}p_m^T p_n+\sum_{\mathcal{F}_k\in\mathcal{O}_m}p_m^T p_k+\left(|\mathcal{C}_m|-|\mathcal{O}_m|\right)\left(\sum_{\mathcal{F}_q\in S_{\mathcal{I}}}\frac{p_m^T p_q}{|S_{\mathcal{I}}|}+\log|S_{\mathcal{I}}|\right)\\ &\approx -\sum_{\mathcal{F}_n\in\mathcal{C}_m}p_m^T p_n+\sum_{\mathcal{F}_k\in\mathcal{O}_m}p_m^T p_k+\left(|\mathcal{C}_m|-|\mathcal{O}_m|\right)\left(\sum_{\mathcal{F}_q\in\mathcal{O}_m}\frac{p_m^T p_q}{|\mathcal{O}_m|}+\log|S_{\mathcal{I}}|\right)\\ &=-\sum_{\mathcal{F}_n\in\mathcal{C}_m}p_m^T p_n+\frac{|\mathcal{C}_m|}{|\mathcal{O}_m|}\sum_{\mathcal{F}_k\in\mathcal{O}_m}p_m^T p_k+\left(|\mathcal{C}_m|-|\mathcal{O}_m|\right)\log|S_{\mathcal{I}}|\\ &=\overline{\psi}(\mathcal{C}_m,\mathcal{O}_m),\end{aligned} \tag{14}$$

where the first inequality is hold by obeying the Jensen's inequality, as the logarithmic function $\log(\cdot)$ is concave.

Since we have $S_{\mathcal{I}} \approx \mathcal{C}_m \cup \mathcal{O}_m$ and $|\mathcal{O}_m| \gg |\mathcal{C}_m|$, the third equation is obtained with the assumption that $S_{\mathcal{I}}$ can be approximated by $\mathcal{O}_m$. Finally, considering the whole feature set $S_{\mathcal{I}}$, our clustering-inspired loss is defined as follows:

$$\mathcal{L}_{\text{CI}} = \frac{1}{|S_{\mathcal{I}}|} \sum_{\mathcal{F}_m \in S_{\mathcal{I}}} \overline{\psi}(\mathcal{C}_m, \mathcal{O}_m). \tag{15}$$

With the help of $\mathcal{L}_{\text{CI}}$, we are able to perform unsupervised clustering learning for ambiguous regions. Meanwhile, for the remaining regions, we adopt the commonly used BCE Loss for the supervision. In this way, crack segmentation models can escape from the disturbance of those ambiguous labels and learn discriminative features. Finally, our overall loss can be formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{CI}}, \tag{16}$$

where the scalar β is used to balance the two loss functions.

4. Experiments

The essential experimental setup is described in Sec 4.1. To find the most suitable parameters λ, γ, and β in CIRL, we carry out a series of experiments on different parameters in Sec. 4.2. Furthermore, to verify the effectiveness of our proposed method, we not only compare it with the existing crack segmentation models in Sec. 4.3, but also compare it with other state-of-the-art segmentation losses in Sec. 4.4. More importantly, in Sec. 4.5, our ablation study first verifies the advantage of the proposed ambiguity-aware segmentation loss over the existing uncertainty-based methods. Then, we provide evidence that the proposed clustering-inspired loss outperforms the existing unsupervised clustering learning methods for crack segmentation tasks.

4.1. Experimental setup

In this paper, we carry out extensive experiments based on our CrackSeg5k dataset, two public crack segmentation datasets, one blood vessel segmentation dataset, and the corresponding implementation details.

CrackSeg5k. Our CrackSeg5k dataset is collected from the nuclear power plants by a high-resolution camera, comprising 2000 images with the size of 7360 $\times$ 4912. The width of the crack in the collected images varies from 0.05 mm to 15 mm. Moreover, the collected images contain different kinds of noise, such as various concrete types and light intensity. To extend this dataset without compromising the resolution and the ratio of the different classes, we directly slice these images into 512 $\times$ 512 pixels, constructing a final dataset with 5560 samples. By following the settings in [23], This dataset is divided into training, validation, and testing sets with 90%, 5%, and 5% ratios.

Crack500 [61]. There are 3368 images in the Crack500 dataset that contain crack images with various shapes and cluttered backgrounds. We should note that this dataset involve a large crack width range and low contrast between crack and non-crack regions. Here, the number of training images, validation images and testing images in this dataset is 1896, 348 and 1124, respectively.

CrackTree200 [66]. This dataset contains 206 images that captured from asphalt pavement. This dataset also suffer from low contrast between cracks and the surrounding pavement. By following previous work [8], we use 164 images for training and the rest for testing in our experiments.

DRIVE [37]. To further verify the scalability of our method, we also use DRIVE dataset, which is designed for blood vessel segmentation on medical segmentation tasks. This dataset only contains 40 retina images. The limited number of training images can evaluate our model's performance on the small dataset. By following previous works [8, 52], we set 20 images for training and 20 for testing.

Implementation details. We carry out our experiments in PyTorch[2] with a single NVIDIA RTX 3090. By following previous work [8], we adopt horizontal flipping, random cropping, and random rotation with 90°, 180° and 270° as our data augmentation strategies. Also, we have the same settings as in [8] that all training samples are cropped to 256 $\times$ 256 during the training. Adam [21] is chosen as our optimizer coupling with an initial learning rate of 10^{-3}, a weight decay of 5×10^{-4}, and a mini-batch size of 2. For these four datasets, the models are trained with 2000 epochs in total. In addition, during the initial period of the training epochs, the learned variances are less informative when the Std head is not properly learned. Thus, the initial 1000 epochs are the first stage, after which our model adopts the joint supervision of $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{CI}}$ (ref to Eq. 16).

Evaluation metrics. To evaluate the pixel-wise accuracy for crack segmentation, we follow the existing works [8, 19, 47] and use F1 score, Precision, and Recall as our metrics. It is worth noting that Precision and Recall are computed by comparing predicted and ground-truth masks at the pixel level. F1 score is computed as: $F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$.

4.2. Sensitivity study on hyper-parameters

Four hyper-parameters have been introduced in this paper. λ is used to help the model to capture ambiguity. γ is used to localize ambiguous regions in the first phase, K is used to define the size of the close neighbor set, and β is exploited to balance the BCE Loss and CI Loss in the second phase. The hyper-parameter sensitivity study on the CrackSeg5k dataset with JTFN [8] as our base segmentation model is introduced in Fig. 4. It can be observed the four hyper-parameters λ, γ, K, and β across a wide range only have 1.5%, 0.8%, 2.0%, 1.8% F1 score decreases compared with the highest, respectively, which demonstrates the potential of our proposed method on real-world applications. Based

[2] https://pytorch.org/

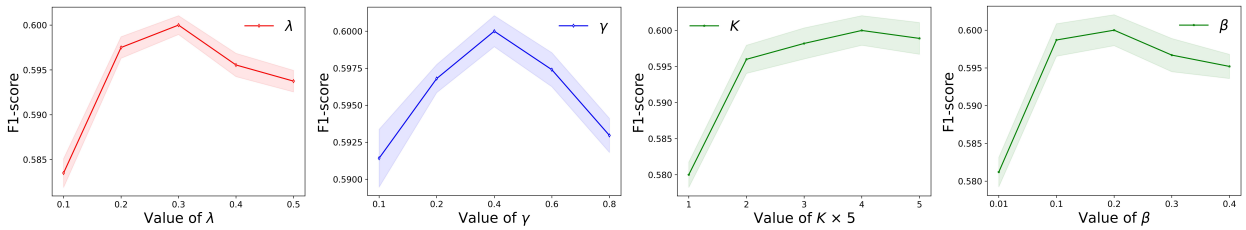

Figure 4. Hyper-parameter sensitivity study of λ, γ, K, and β on CrackSeg5k dataset.

Method	Venue	CrackTree200			Crack500			DRIVE			CrackSeg5K		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
UNet [46]	MICCAI	79.16	78.95	78.42	62.22	68.85	61.83	82.74	80.59	81.41	61.49	57.19	56.28
VGG-UNet [40]	CVPR	83.49	80.43	81.84	58.18	60.26	51.79	81.17	82.05	81.25	63.21	56.83	55.92
TopoNet [19]	NeurIPS	81.85	77.80	79.03	66.81	62.68	60.06	82.94	80.29	81.36	61.39	56.25	57.72
DRU [55]	ICCV	84.80	77.46	80.49	61.94	71.43	62.82	84.36	80.82	82.30	63.27	58.28	58.15
Crackformer [32]	ICCV	84.13	81.93	83.42	69.13	66.24	64.75	83.05	81.19	83.25	66.29	59.19	58.83
JTFN [8]	ICCV	85.87	82.58	84.19	68.81	69.06	65.76	82.71	83.40	82.81	65.14	58.33	58.42
JTFN + CIRL (Our)	-	**87.62**	**83.92**	**86.53**	**70.32**	**69.93**	**67.62**	**84.57**	**82.95**	**84.32**	**67.37**	**59.46**	**60.08**

Table 2. Comparisons with the state-of-the-art crack segmentation methods on three crack datasets and one blood vessel dataset.

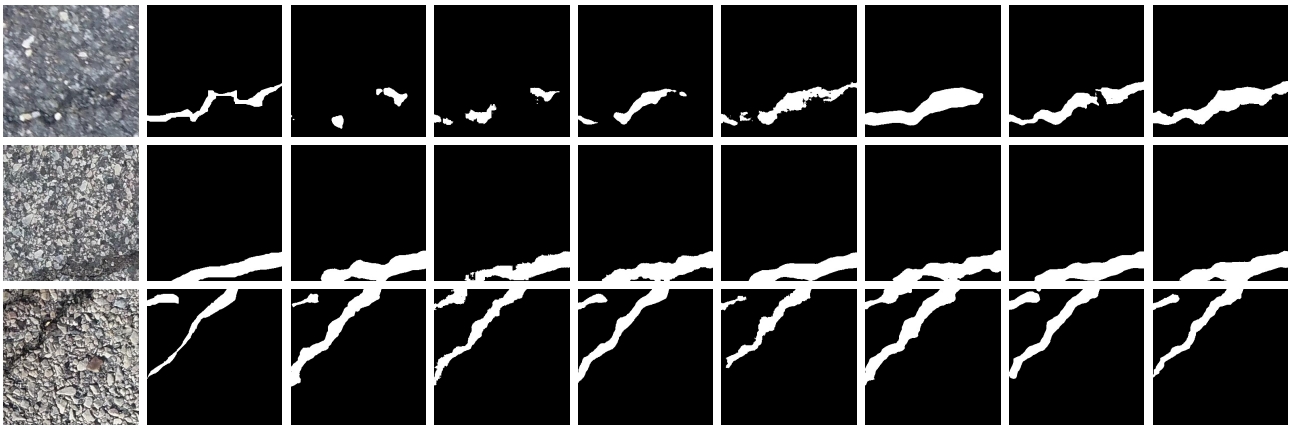

Figure 5. Demonstration of segmentation results comparisons on Crack500 dataset. From left to right: Input image, ground-truth masks, results of the UNet, VGG-UNet, TopoNet, DRU, Crackformer, JTFN, and our results.

on these observations, we set $\lambda = 0.3$, $\gamma = 0.4$, $K = 4 \times 5$, and $\beta = 0.2$ in our next experiments.

4.3. Comparison on crack segmentation models

In this section, our method is compared to several baselines including state-of-the-art segmentation methods. As shown in Tab 2, when compared with the original JTFN, CIRL obtains performance gain around 2.34 % and 1.86 % with F1 score on the CrackTree200 and Crack500 dataset. From these results, we could infer that the proposed CIRL can work together with the boundary-preserving methods, without sacrificing the gain of the supervision for the crack boundary. Moreover, Fig. 5 shows segmentation examples of CIRL and the other methods on the Crack500 dataset. We can see that CIRL delineates cracks better compared with alternatives.

4.4. Comparison on segmentation losses

To further validate our method, we conduct a series of experiments on the CrackSeg5k dataset and report quantitative results with Precision, Recall, and F1 scores to verify the effectiveness of CIRL. Herein, we adopt the JTFN [8] as the base segmentation model. As the ultimate goal of CIRL is to enhance the model's ability to extract discriminative features, we also compare it with other loss functions that have the same intuition: including Focal Loss [30], Robust Dice Loss [53], clDice Loss [48], Poly Loss [25], and Robust T-Loss [15]. Tab. 3 shows the results compared with

the existing loss functions. We can observe that CIRL outperforms BCE Loss by a large margin. This benefits from that the learned variances allow us to localize the ambiguous regions. Then, CIRL can help the model escape from the disturbance of those ambiguous labels by starting from an unsupervised clustering manner for the above regions.

Method	Venue	Precision	Recall	F1
Focal Loss [30]	ICCV	65.71	58.94	58.85
Robust Dice Loss [53]	TMI	64.23	57.83	59.06
clDice Loss [48]	CVPR	64.31	58.49	59.32
Poly Loss [25]	ICLR	66.14	57.35	59.56
Robust T-Loss [15]	MICCAI	65.72	58.14	58.84
CIRL (Our)	-	**67.37**	**59.46**	**60.08**

Table 3. Comparisons of different loss functions on the CrackSeg5k with JTFN as the base segmentation model.

4.5. Ablation study

To examine the contribution of each component in our proposed framework: Aseg Loss in Phase I and CI Loss in Phase II, extensive experiments are performed on CrackSeg5k with JTFN [8] as the base segmentation model.

Evaluation on different phases To examine the contribution of each phase in our proposed framework, we provide a series of experiments on CrackSeg5k with JTFN as the base segmentation model.

Method	Precision	Recall	F1
PhaseI	65.17	58.78	58.57
PhaseI + PhaseII	**67.37**	**59.46**	**60.08**

Table 4. Evaluation on different phases.

Table 4 shows that phase II refreshes the performance achieved by phase I. The reason behind this effect is twofold: (i) In phase I, the proposed Aseg Loss allows us to localize the ambiguous regions from marginal non-crack regions. (ii) Then, our phase II is able to alter the existing supervised learning for the above ambiguous regions in an unsupervised clustering manner, and learn discriminative features with the help of CI Loss.

Evaluation on the CI Loss in Phase II. As expected, JTFN [8] achieves better results by combining the CI Loss. Table 5 shows that it improves the precision, recall, and F1 score by 2.2%, 0.68%, and 1.51 over the current clustering learning-based method, respectively. The reason behind this effect is that our CI Loss performs contrastive clustering learning between two feature sets, instead of positive and negative pairs. In this regard, our method can alleviate the class collision issue in the existing contrastive learning. Moreover, Fig. 6 demonstrates that our loss helps the model to learn discriminative features for marginal non-crack regions.

Method	Venue	Precision	Recall	F1
NNCLR [13]	ICCV	65.45	58.91	58.95
CI Loss (Our)	-	**67.37**	**59.46**	**60.08**

Table 5. Performance comparisons of CI Loss and NNCLR.

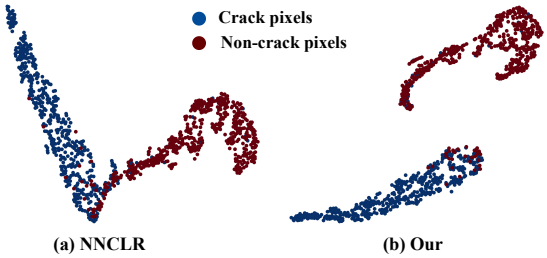

Figure 6. T-SNE visualization of the features in marginal regions.

5. Conclusion

In summary, the ambiguities in marginal non-crack regions can hinder the performance of state-of-the-art crack segmentation models. In this paper, a two-phase clustering-inspired representation learning (CIRL) framework is proposed for learning more accurate pixel-level crack segmentation. The CIRL contains the ambiguity-aware segmentation loss (Aseg Loss) that enforces the network learns to predict segmentation variance for each pixel. The resulting variances allow us to localize the ambiguous region further. Then, a clustering-inspired loss (CI Loss) is proposed for learning the discriminative features of the above regions. The experiments on crack segmentation tasks demonstrate that our method outperforms state-of-the-art crack segmentation approaches and loss functions. Since previous methods shows great power in a frequency domain for learning discriminative features [39, 57], we plan to further consider the feature-based data augmentation [6] and neuron-based spiking network [27, 28] for crack segmentation tasks.

Acknowledgements. This work is supported by the National Natural Science Funds for Distinguished Young Scholar under Grant 62325307, the National Nature Science Foundation of China under Grants 62073225, 62203134, 62072315, the National Key R&D Program of China under Grants 2020YFA0908700, the Natural Science Foundation of Guangdong Province under Grants 2023B1515120038, Shenzhen Science and Technology Innovation Commission (20220809141216003, JCYJ20210324093808021, JCYJ20220531102817040), the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, the Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235, the Scientific Instrument Developing Project of Shenzhen University under Grant 2023YQ019.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 2, 4

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1, 2

[3] Priti S Chakurkar, Deepali Vora, Shruti Patil, Sashikala Mishra, and Ketan Kotecha. Data-driven approach for ai-based crack detection: Techniques, challenges, and future scope. *Front. Sustain. Cities*, 5:1253627, 2023. 1

[4] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *ICCV*, pages 5879–5887, 2017. 3

[5] Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Jie Chen, Zun Liu, and Jianqiang Li. Geometry-aware guided loss for deep crack recognition. In *CVPR*, pages 4703–4712, 2022. 1

[6] Zhuangzhuang Chen, Jin Zhang, Pan Wang, Jie Chen, and Jianqiang Li. When active learning meets implicit semantic data augmentation. In *ECCV*, pages 56–72, 2022. 8

[7] Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Guanming Zhu, Zun Liu, Jie Chen, and Jianqiang Li. The devil is in the crack orientation: A new perspective for crack detection. In *ICCV*, pages 6653–6663, 2023. 1

[8] Mingfei Cheng, Kaili Zhao, Xuhong Guo, Yajing Xu, and Jun Guo. Joint topology-preserving and feature-refinement network for curvilinear structure segmentation. In *ICCV*, pages 7147–7156, 2021. 1, 2, 6, 7, 8

[9] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *ICCV*, pages 1142–1160, 2022. 3

[10] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *CVPR*, pages 13693–13702, 2021. 2, 3

[11] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005. 3

[12] Quang Du Nguyen and Huu-Tai Thai. Crack segmentation of imbalanced data: The role of loss functions. *Engineering Structures*, 297:116988, 2023. 2

[13] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, pages 9588–9597, 2021. 3, 8

[14] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. *NeurIPS*, 17, 2004. 5

[15] Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Gröger, Marc Pouly, and Alexander A Navarini. Robust t-loss for medical image segmentation. In *MICCAI*, pages 714–724, 2023. 7, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[17] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, pages 2888–2897, 2019. 4

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[19] Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *NeurIPS*, 32, 2019. 6, 7

[20] Wenlian Huang, Guanming Zhu, Qiang Huang, Zhuangzhuang Chen, Jie Chen, and Jianqiang Li. Defect screening on nuclear power plant concrete structures: A two-staged method based on contrastive representation learning. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 691–697, 2023. 1

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Jacob König, Mark Jenkins, Mike Mannion, Peter Barrie, and Gordon Morison. What's cracking? a review and analysis of deep learning methods for structural crack segmentation, detection and quantification. *arXiv preprint arXiv:2202.03714*, 2022. 1

[23] Shreyas Kulkarni, Shreyas Singh, Dhananjay Balakrishnan, Siddharth Sharma, Saipraneeth Devunuri, and Sai Chowdeswara Rao Korlapati. Crackseg9k: a collection and benchmark for crack segmentation datasets and frameworks. In *European Conference on Computer Vision*, pages 179–195, 2022. 6

[24] Gyumin Lee, Seung Jun Lee, and Changyong Lee. A convolutional neural network model for abnormality diagnosis in a nuclear power plant. *Applied Soft Computing*, 99:106874, 2021. 1

[25] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Jay Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. In *International Conference on Learning Representations*, 2021. 7, 8

[26] Kai Li, Bo Wang, Yingjie Tian, and Zhiquan Qi. Fast and accurate road crack detection based on adaptive cost-sensitive loss function. *IEEE Transactions on Cybernetics*, 53(2): 1051–1062, 2023. 2

[27] Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Xiaopeng Fan, and Yonghong Tian. Neuron-based spiking transmission and reasoning network for robust image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 8

[28] Wenrui Li, Xi-Le Zhao, Zhengyu Ma, Xingtao Wang, Xiaopeng Fan, and Yonghong Tian. Motion-decoupled spiking transformer for audio-visual zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3994–4002, 2023. 8

[29] Jianghai Liao, Yuanhao Yue, Dejin Zhang, Wei Tu, Rui Cao, Qin Zou, and Qingquan Li. Automatic tunnel crack inspection using an efficient mobile imaging module and a lightweight cnn. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15190–15203, 2022. 1, 2

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 7, 8

[31] Gaoyang Liu, Wei Ding, Jiangpeng Shu, Alfred Strauss, Yuanfeng Duan, et al. Two-stream boundary-aware neural network for concrete crack segmentation and quantification. *Structural Control and Health Monitoring*, 2023, 2023. 2

[32] Huajun Liu, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, and Hui Kong. Crackformer: Transformer network for fine-grained crack detection. In *ICCV*, pages 3783–3792, 2021. 1, 2, 3, 7

[33] Jingwei Liu, Xu Yang, Stephen Lau, Xin Wang, Sang Luo, Vincent Cheng-Siong Lee, and Ling Ding. Automated pavement crack detection and segmentation based on two-step convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, 35(11):1291–1305, 2020. 2

[34] Shuo Liu, Wenrui Ding, Chunhui Liu, Yu Liu, Yufeng Wang, and Hongguang Li. Ern: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sensing*, 10(9):1339, 018. 2

[35] Zhenqing Liu, Yiwen Cao, Yize Wang, and Wei Wang. Computer vision-based concrete crack detection using u-net fully convolutional networks. *Automation in Construction*, 104: 129–139, 2019. 2

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[37] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Deep retinal image understanding. In *MICCAI*, 2016. 6

[38] Dimitrios Marmanis, Konrad Schindler, Jan Dirk Wegner, Silvano Galliani, Mihai Datcu, and Uwe Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018. 2

[39] Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using random frequency components. In *ICCV*, pages 19673–19684, 2023. 8

[40] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *CVPR*, pages 3136–3145, 2018. 7

[41] Son Dong Nguyen, Thai Son Tran, Van Phuc Tran, Hyun Jong Lee, Md Jalil Piran, and Van Phuc Le. Deep learning-based crack detection: A survey. *International Journal of Pavement Research and Technology*, 16(4):943–967, 2023. 1, 2

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[43] Fabio Panella, Aldo Lipani, and Jan Boehm. Semantic segmentation of cracks: Data challenges and architecture. *Automation in Construction*, 135:104110, 2022. 2

[44] Bryan G Pantoja-Rosero, D Oner, Mateusz Kozinski, Radhakrishna Achanta, Pascal Fua, Fernando Pérez-Cruz, and K Beyer. Topo-loss for continuity-preserving crack detection using deep learning. *Construction and Building Materials*, 344:128264, 2022. 3

[45] Zhong Qu, Wen Chen, Shi-Yan Wang, Tu-Ming Yi, and Ling Liu. A crack detection algorithm for concrete pavement based on attention mechanism and multi-features fusion. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11710–11719, 2021. 1

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 2, 7

[47] Mojtaba Seyedhosseini, Mehdi Sajjadi, and Tolga Tasdizen. Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks. In *ICCV*, pages 2168–2175, 2013. 6

[48] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *CVPR*, pages 16560–16569, 2021. 7, 8

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017. 1, 2

[52] Feigege Wang, Yue Gu, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Context-aware spatio-recurrent curvilinear structure segmentation. In *CVPR*, pages 12648–12657, 2019. 6

[53] Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang. A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE Transactions on Medical Imaging*, 39(8): 2653–2663, 2020. 7, 8

[54] Sen Wang, Xing Wu, Yinghui Zhang, Xiaoqin Liu, and Lun Zhao. A neural network ensemble method for effective crack segmentation using fully convolutional networks and multi-scale structured forests. *Machine Vision and Applications*, 31:1–18, 2020. 2

[55] Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Recurrent u-net for resource-constrained segmentation. In *ICCV*, pages 2142–2151, 2019. 7

[56] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *ICCV*, pages 8150–8159, 2019. 3

[57] Qinwei Xu, Ruipeng Zhang, Ziqing Fan, Yanfeng Wang, Yi-Yan Wu, and Ya Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023. 8

[58] Yang Xu, Yunlei Fan, and Hui Li. Lightweight semantic segmentation of complex structural damage recognition

for actual bridges. *Structural Health Monitoring*, page 14759217221147015, 2023. 2

[59] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *NeurIPS*, pages 5802–5815, 2022. 2, 5

[60] Xincong Yang, Heng Li, Yantao Yu, Xiaochun Luo, Ting Huang, and Xu Yang. Automatic pixel-level crack detection and measurement using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering*, 33 (12):1090–1109, 2018. 2

[61] Lei Zhang, Fan Yang, Yimin Daniel Zhang, and Ying Julie Zhu. Road crack detection using deep convolutional neural network. In *ICIP*, pages 3708–3712, 2016. 6

[62] Yuhui Zhang, Yuichiro Wada, Hiroki Waida, Kaito Goto, Yusaku Hino, and Takafumi Kanamori. Deep clustering with a constraint for topological invariance based on symmetric infonce. *Neural Computation*, 35:1288–1339, 2023. 3

[63] Xiaoyu Zhao, Wenlian Huang, Jie Chen, Zhuangzhuang Chen, and Jianqiang Li. Automatic thin crack segmentation with deep context aggregation network. In *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 206–212, 2022. 1

[64] Quan Zhou, Yong Qiang, Yuwei Mo, Xiaofu Wu, and Longin Jan Latecki. Banet: Boundary-assistant encoder-decoder network for semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25259–25270, 2022. 2

[65] Shanglian Zhou, Carlos Canchila, and Wei Song. Deep learning-based crack segmentation for civil infrastructure: data types, architectures, and benchmarked performance. *Automation in Construction*, 146:104678, 2023. 2

[66] Qin Zou, Yu Cao, Qingquan Li, Qingzhou Mao, and Song Wang. Cracktree: Automatic crack detection from pavement images. *Pattern Recognition Letters*, 33(3):227–238, 2012. 6

[67] Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE transactions on image processing*, 28(3):1498–1512, 2018. 1, 2