SemiCurv: Semi-Supervised Curvilinear Structure Segmentation

Xun Xu, *Senior Member, IEEE*, Manh Cuong Nguyen, Yasin Yazici, Kangkang Lu, Hlaing Min, and Chuan-Sheng Foo

***Abstract*— Recent work on curvilinear structure segmentation has mostly focused on backbone network design and loss engineering. The challenge of collecting labelled data, an expensive and labor intensive process, has been overlooked. While labelled data is expensive to obtain, unlabelled data is often readily available. In this work, we propose SemiCurv, a semi-supervised learning (SSL) framework for curvilinear structure segmentation that is able to utilize such unlabelled data to reduce the labelling burden. Our framework addresses two key challenges in formulating curvilinear segmentation in a semi-supervised manner. First, to fully exploit the power of consistency based SSL, we introduce a geometric transformation as strong data augmentation and then align segmentation predictions via a differentiable inverse transformation to enable the computation of pixel-wise consistency. Second, the traditional mean square error (MSE) on unlabelled data is prone to collapsed predictions and this issue exacerbates with severe class imbalance (significantly more background pixels). We propose a N-pair consistency loss to avoid trivial predictions on unlabelled data. We evaluate SemiCurv on six curvilinear segmentation datasets, and find that with no more than 5% of the labelled data, it achieves close to 95% of the performance relative to its fully supervised counterpart.**

***Index Terms*— Semi-supervised learning, semantic segmentation.**

I. Introduction

CURVILINEAR structure segmentation aims to extract thin, curvilinear structures from images. It has wide applications, including road crack segmentation, road segmentation from satellite images and biomedical image segmentation (e.g. segmenting blood vessels and cells). Existing research into curvilinear structure segmentation has focused on designing better network architectures [1], [2] or loss functions (e.g. topology loss [3], [4]) to better account for the specific nature of curvilinear structures like connectivity and loop formation.

A major challenge in curvilinear structure segmentation that has received less attention is the need for a large amount of labelled data to obtain good performance. In many cases, acquiring such labelled data for segmentation tasks is non-trivial and sometimes requires expert labelling. Moreover, the fine detail present in curvilinear structures demands increased annotation effort in comparison to the coarse-grained polygon-based annotation that is sufficient for general (object-based) segmentation tasks. At the same time, we often have access to unlabelled data at very low cost. For example, road scanning vehicles are running everyday to collect road surface photographs, satellites are collecting aerial images on a daily basis, and medical images are abundant from daily diagnoses. This availability of unlablled data suggests the use of semi-supervised learning (SSL) approaches that are able to exploit the unlabelled data to reduce the amount of labelled data required.

Much work in SSL has focused on semi-supervised classification [5], and far fewer works have studied the use of SSL for image segmentation, much less curvilinear structure segmentation. Existing SSL methods for semantic segmentation adopt one of two approaches: 1) utilizing a Generative Adversarial Network (GAN) to enforce that predicted segmentation maps are indistinguishable from real segmentation maps [6], [7] or 2) enforcing consistency of predicted segmentation maps on unlabelled data between two augmented samples [8]–[10]. The first strategy still requires a considerable amount of labelled data to train the GAN. In contrast, consistency-based approaches are able to work with much less labelled data and are more promising in the low-label regime. Therefore, we build upon the consistency-based approach towards semi-supervised segmentation in this work.

Even though previous work has developed methods for semi-supervised segmentation, we argue that direct application of these methods to the curvilinear segmentation task is sub-optimal for the following two reasons. First, the types of data augmentation used by state-of-the-art SSL methods for generic segmentation tasks has thus far been limited to pixel-wise perturbations, including a mix-up of multiple images [9], [10]. Pixel-wise perturbations do not create more diverse geometric views of image data (e.g. vertical cracks will always remain vertical under pixel-wise perturbations), and thus may not fully exploit the power of consistency based SSL. Moreover, mixing up images could shift the data away from the real data manifold (i.e. the mixed-up image may not look real), potentially harming SSL as demonstrated in our experiments. Secondly, curvilinear structures typically

Manuscript received 26 September 2021; revised 23 April 2022; accepted 9 May 2022. Date of publication 27 July 2022; date of current version 3 August 2022. This work was supported in part by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds under Grant A20H6b0151 and in part by the National Natural Science Foundation of China under Grant 6210021203. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhao Zhang. *(Corresponding author: Xun Xu.)*

The authors are with the Institute for Infocomm Research (I2R), A*STAR, Singapore 138632 (e-mail: xux@i2r.a-star.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3189823, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3189823

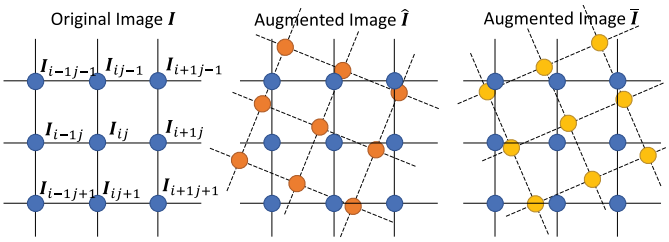


Fig. 1. Illustration of aligning augmented images. Two random augmentations are applied to original image $\mathbf{I}$, resulting in $\hat{\mathbf{I}}$ and $\bar{\mathbf{I}}$. As pixel-to-pixel correspondence is no longer available between $\hat{\mathbf{I}}$ and $\bar{\mathbf{I}}$, we apply additional inverse transformation to align two augmented images.

occupy a small fraction (often under 10% and as little as 1% in the datasets we considered) of the image compared to the background, leading to a severely class-imbalanced segmentation target; this is unlike the more general semantic segmentation datasets [11] considered by these existing methods. As we will show, the class imbalance and sparsity leads to a "collapsing" issue with the consistency loss used by SSL methods, resulting in models predicting constant outputs.

We next describe the key improvements we make to consistency-based SSL to address these challenges associated specifically with curvilinear segmentation tasks:

1) Differentiable affine transformation. First, we propose affine transformations as data augmentation for curvilinear images. This augmentation is effective because of two reasons. First, images with curvilinear structures are often captured from planar surfaces; this is characteristic of images of road surfaces, satellite images, and images of cells taken under a microscope, to give a few examples. Hence, synthesizing novel viewpoints as augmentation by affine transformation is valid [12]. Second, segmentation of curvilinear structures is expected to be affine equivariant: for instance, the notion of a crack on a road surface should remain the same regardless of rotation, translation or resizing of the image, unlike in semantic segmentation tasks where the pose of an semantic object is often fixed or subject to small variation.

One challenge with using affine transformations as data augmentation is that despite being more diverse, they prohibit computing consistency loss at the pixel level between two randomly augmented images because pixel-wise correspondence is not preserved during affine transformation. As illustrated in Fig. 1, supposing $\hat{\mathbf{I}}$ and $\bar{\mathbf{I}}$ are randomly augmented from $\mathbf{I}$, there is no explicit pixel-wise correspondence between $\hat{\mathbf{I}}$ and $\bar{\mathbf{I}}$ with which we can enforce the predictions to be consistent. To address this alignment issue, we propose to employ an inverse affine transformation following the networks' predictions as illustrated in Fig. 2. With both forward and inverse transformations, the pixel-to-pixel correspondence is restored, enabling the computation of the pixel-wise consistency loss between student and teacher models. We note that incorporating the inverse transformation into the networks requires it to be differentiable for gradients to backpropagate; as affine transformation is differentiable, the inverse transformation allows end-to-end learning as part of the networks.

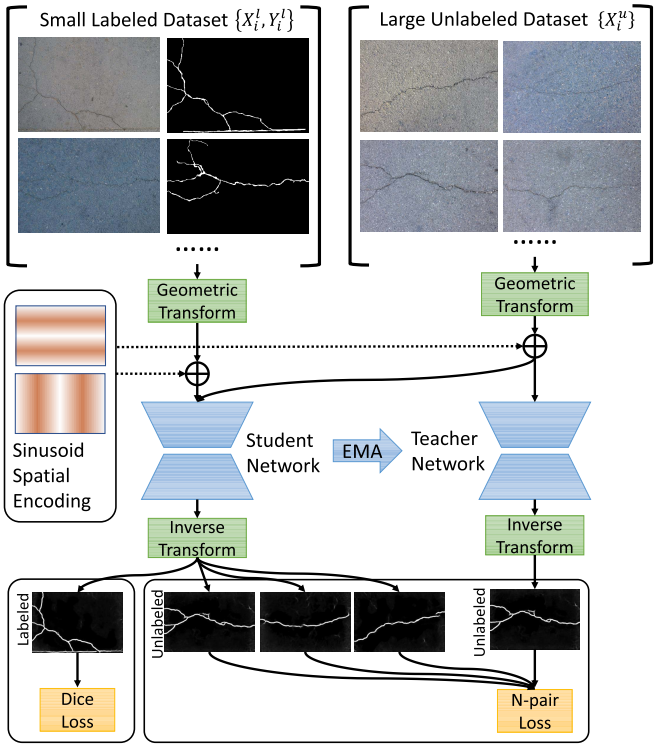


Fig. 2. Overview of SemiCurv, a semi-supervised learning framework for curvilinear structure segmentation. We adopt a mean-teacher like architecture for semi-supervised learning. i) We first propose a geometric transformation and the inverse transformation to allow stronger data augmentation. ii) To account for the severe class-imbalance issue, we employ Dice loss and N-Pair loss for labelled and unlabelled data respectively. iii) Spatial encoding is further incorporated to capture spatial correlations.

2) N-pair consistency. As noted above, we observe that the widely adopted consistency loss, mean square error (MSE), is prone to a "collapsing" issue. Trivial solutions to minimize the MSE on unlabelled data exist where both student and teacher networks predict constant outputs. The introduction of a non-learnable teacher model may partially address this issue, but cannot guarantee in principle that a trivial solution is not learned. Moreover, we notice that there are often far fewer positive (foreground) pixels than negative ones (background) in curvilinear structure datasets, suggesting the prior distribution is highly biased towards background. This imbalance further encourages the trivial solution where the learned model predicts all background on the unlabelled data. To overcome this challenge, we propose to use an N-pair loss [13] to avoid trivial predictions on unlabelled data, inspired by the recent success of contrastive learning [14], [15].

3) Spatial coordinate encoding. Finally, there is clear spatial connectivity and correlation in curvilinear structure patterns, in that positive pixels are spatially adjacent. Although topology was studied to improve connectivity [3], [4], they require fully labelled ground-truth to provide topological supervision. To exploit the spatial correlation prior, we propose to add spatial coordinate encoding as features to all layers of the backbone segmentation network. To avoid the potential overfitting to absolute spatial coordinates [16], we encode spatial adjacency using a sinusoid function, which preserves spatial adjacency information while reducing risk of overfitting.

In summary, our work makes the following contributions:

- We introduce a differentiable affine transformation to fully exploit the strength of consistency based semi-supervised learning for curvilinear image segmentation.
- We provide insight into a "collapsing" issue with MSE consistency loss defined on unlabelled data and propose to use N-pair loss to mitigate this issue.
- To further capture the spatial correlation, we incorporate sinusoid spatial encoding as additional features.
- To the best of our knowledge, this is the first attempt to investigate into curvilinear structure segmentation in a semi-supervised fashion. We extensively benchmarked state-of-the-art semi-supervised methods on six curvilinear segmentation datasets.

II. Related Work

A. Curvilinear Structure Segmentation

We briefly review previous work on curvilinear structure segmentation in the context of road segmentation from satellite images, medical image segmentation and crack detection in pavements. Early work in road segmentation [17] used fully connected networks, but were soon followed by CNNs [18]. Recent work has built on the CNN approach, focusing on designing deeper and wider networks and considering information at multiple scales [19], [20]. In the medical imaging community, efforts have focused on further improving the popular UNet backbone [21], [22]. These fully supervised methods have shown to be competitive in delineating linear structures like membrane segmentation [23]. Progress in the area of crack segmentation proceeded in a similar trajectory, where recent progress is due to improved backbone network design [1], [24]. Orthogonal to the above, in this work, we address the semi-supervised learning setting to reduce the amount of labelled data needed to train high performance models by exploiting unlabelled data.

To model spatial correlation and connectivity in curvilinear structures, a few works have incorporated topological constraints when training the segmentation model. One study utilized intermediate features from a VGG network to capture topological structure [4], while another work used concepts from persistent homology to provide more principled topological features; this idea was implemented by defining a loss between ground-truth and predicted persistent diagrams [3], which assumes access to ground-truth labels. In this work, we adopt a simpler approach by using positional encodings to capture spatial correlation.

B. Semi-Supervised Learning

We briefly review the two genres of semi-supervised learning (SSL) methods, namely the consistency-based SSL and adversarial training based SSL; for a more complete review, please refer to [5]. Consistency-based SSL methods [25]–[27] utilize the unlabelled data to constrain learning by enforcing the model's predictions on the unlabelled data to be consistent with a specified target. Consistency regularization was firstly introduced in the context of deep learning by [28]. To better exploit the power of ensemble learning [25] used temporal ensembling of models' predictions as a target for consistency. [26] further proposed to ensemble the model parameters over time to obtain more stable consistency targets. Different from the previously mentioned approaches, the VAT method [27] proposed to enforce consistency of a model's predictions with that on an adversarially perturbed input [29]. Orthogonal to the above, the MixMatch method [30] proposed to interpolate between labelled and unlabelled samples to diversify the training set. Due to the high accuracy of temporal ensembling and low memory footprint, we adopted Mean Teacher [26] as the backbone framework. Another line of works adapt generative adversarial training to semi-supervised learning. Specifically, in the seminal work [31], a discriminator was introduced to distinguish generated samples from real ones and at the same time classify real samples into respective categories.

C. Semi-Supervised Segmentation

Semi-Supervised learning has been applied to semantic image segmentation to alleviate the expensive image annotation task. Following the adversarial training approach [31], [32] first employed a discriminator to distinguish generated fake images from real ones to improve semi-supervised learning. A conditional generative adversarial network (GAN) was further employed by ensuring that a discriminator cannot distinguish predicted segmentation masks on unlabelled images from ground-truth segmentation masks (on the labelled images) [6]; [7] further imposes cycle consistency (building on CycleGAN) to constrain the learning on unlabelled data. However, under low labeled data training a GAN is prone to instability and may not generalize well to unlabeled data. Following the consistency-based SSL approaches, [9] proposed to enforce consistency between the segmentation predictions of images subject to a MixUp augmentation, i.e. random crops of the two images are superimposed and consistency is enforced on the segmentation predictions of respective patches. In the context of medical image segmentation, a framework similar to Mean Teacher has been applied to skin lesion and CT-scan segmentation [33]. All these existing methods only consider data augmentations that are either pixel-wise perturbations [9] or rotations that are multiples of 90° [33] with pairwise consistency; such augmentations are not strong enough to fully exploit the power of consistency regularization. These works also do not address the issue where the MSE loss used to enforce consistency is prone to collapse.

III. Methodology

A. Overview of SemiCurv

We first provide an overview of our SemiCurv approach, then describe the individual components in greater detail. Our SemiCurv approach, illustrated in Fig. 2, is built upon the consistency-based SSL framework (Sect. III-B). Both labelled and unlabelled images are first augmented by a differentiable geometric transformation (Sect. III-C) and then fed into student and teacher networks. The teacher network is a parameter-wise temporal ensemble (exponential moving average) of the student network and generates a target (pseudo-label) to encourage consistency on

the unlabelled data. Sinusoid positional encoding is appended to each convolution layer in the encoder to explicitly capture spatial correlations (Sect. III-E). The predicted segmentation maps (posterior probabilities) from both networks are transformed back to the original pose via the differentiable inverse transform, and the dice loss (Sect. III-F) and N-pair loss (Sect. III-D) are used on the labelled and unlabelled data respectively for training; in particular, we show how the N-pair loss resolves a "collapsing" issue associated with the MSE loss commonly used for consistency on the unlabelled data.

B. Consistency-Based Semi-Supervised Learning

We denote a training set of N_l labelled images and their ground-truth segmentation maps as $\mathcal{D}_l = \{(\mathbf{X}_i^l, \mathbf{Y}_i^l)\}_{i=1}^{N_l}$, and denote the set of N_u unlabelled images as $\mathcal{D}_u = \{\mathbf{X}_j^u\}_{j=1}^{N_u}$. The loss function optimized by semi-supervised learning methods can be generally written as Eq. (1), where l_l, l_u, and γ are supervised loss on labelled images, unsupervised loss on unlabelled images and balancing hyperparameter, respectively.

$$L(\mathcal{D}_l, \mathcal{D}_u) = \frac{1}{N_l}\sum_{i=1}^{N_l} l_l(\mathbf{X}_i^l, \mathbf{Y}_i^l) + \gamma \frac{1}{N_u}\sum_{j=1}^{N_u} l_u(\mathbf{X}_j^u) \quad (1)$$

We note that the loss functions of multiple SSL methods [25]–[27] can be written in this form. Most existing works adopt the cross-entropy loss or its variants as the supervised loss l_l for classification tasks. In this work, we demonstrate in Sect. III-F that for segmentation with severe class imbalance, Dice loss is more appropriate. For the unsupervised loss, the Π model [25], mean teacher model (MT) [26], VAT [27], MixMatch [30] and variants, all apply mean square error (MSE) consistency between the prediction posteriors of an unlabelled image under two different augmentations $t_1(\cdot)$ and $t_2(\cdot)$, and two different segmentation networks $f_1(\cdot)$ and $f_2(\cdot)$ as in Eq. 2. We show in Sect. III-D that under severe class imbalance the N-pair loss is more effective than MSE in avoiding trivial solutions.

$$l_u(\mathbf{X}) = ||f_1(t_1(\mathbf{X})) - f_2(t_2(\mathbf{X}))||_F^2 \quad (2)$$

SemiCurv is based on the mean teacher (MT) method [26] for semi-supervised classification, which has also been previously applied to the segmentation task. Briefly, the MT method maintains two networks: one fully trainable network called the student network, and another non-trainable network called the teacher network. The teacher network provides pseudo-labels to supervised the student network to learn. To improve stability of training, the parameters of teacher are the exponential moving average (EMA) of the student network's parameters. In the rest of the paper, we denote the student network as $f(\cdot)$ and teacher network as $\hat{f}(\cdot)$.

C. Differentiable Geometric Transformation for Augmentation

Strong data augmentation is key to the success of consistency-based semi-supervised learning [34]. As discussed, existing semi-supervised segmentation methods often consider a limited set of augmentations. To further increase the variation of poses to improve feature learning we introduce affine transformations for stronger data augmentation. We denote the transformation applied to input image $\mathbf{X}$ as $t(\mathbf{X})$. To enable computing pixel-wise consistency loss, we further apply the inverse transformation to the output of both student and teacher networks as $t^{-1}(f(t(\mathbf{X})))$. With the inverse transformation, predictions on two arbitrarily augmented images are directly comparable at the pixel level.

Concretely, an affine transformation involves an arbitrary combination of scaling, rotation, shearing and translation, and can be formulated as a matrix multiply with the following transformation matrix,

$$\mathbf{H} = \begin{bmatrix} a_{11} & a_{12} & d_x \\ a_{21} & a_{22} & d_y \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

For every pixel at $(\hat{w}, \hat{h})$ in the image after affine transformation, we find its coordinate in the original image by $[w, h, 1]^\top = \mathbf{H}^{-1}[\hat{w}, \hat{h}, 1]^\top$. We employ bilinear interpolation to produce the pixel intensity at every pixel after transformation. Concretely, we can find the 4 neighbouring pixels before the transform as $\{x_{\lfloor w \rfloor, \lfloor h \rfloor}, x_{\lceil w \rceil, \lfloor h \rfloor}, x_{\lfloor w \rfloor, \lceil h \rceil}, x_{\lceil w \rceil, \lceil h \rceil}\}$. The pixel intensity value after transformation $\hat{x}_{\hat{w},\hat{h}}$ can be determined by bilinear interpolation as,

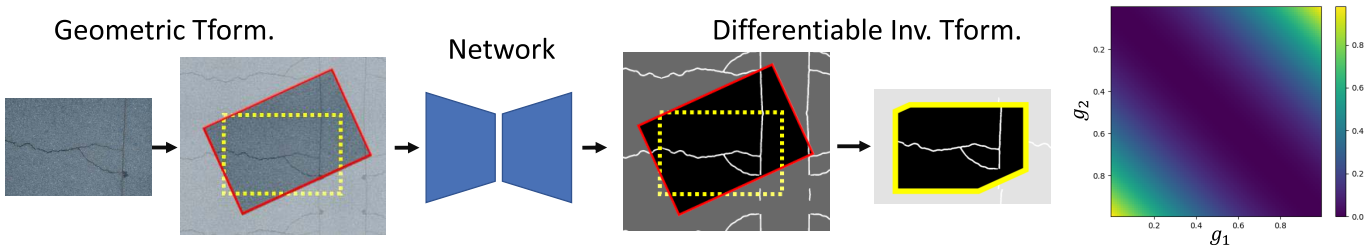
$$\hat{x}_{\hat{w},\hat{h}} = [\lceil w \rceil - w, w - \lfloor w \rfloor] \begin{bmatrix} x_{\lfloor w \rfloor, \lfloor h \rfloor} & x_{\lceil w \rceil, \lfloor h \rfloor} \\ x_{\lfloor w \rfloor, \lceil h \rceil} & x_{\lceil w \rceil, \lceil h \rceil} \end{bmatrix} \begin{bmatrix} \lceil h \rceil - h \\ h - \lfloor h \rfloor \end{bmatrix} \quad (4)$$

Because of the bilinear transformation $\hat{x}_{\hat{w},\hat{h}}$ is obviously differentiable w.r.t. $\{x_{w,h} | w \in \{0, W-1\}, h \in \{0, H-1\}$ where H and W are the height and width of image respectively, i.e. $t(\mathbf{X})$ is differentiable w.r.t. $\mathbf{X}$. For the inverse transformation, we simply apply $\mathbf{H}^{-1}$ for affine transformation and thus $t^{-1}(f(t(\mathbf{X})))$ is still differentiable w.r.t. $\mathbf{X}$.

We also notice that to generate diverse and realistic augmented image samples, we extrapolate images by mirroring the image over the edges. To give an example, the result of a 1D sequence "$abcd$" after mirror flipping is "$dcb|abcd|cba$". This is achieved by clipping the pixel location before transformation with $x = (x//W\%2)*(W - x\%W) + (1 - x//W\%2)*(x\%W)$ where // and % are floor and modulo divisions respectively. An illustration of augmentation for image segmentation is given in Fig. 3 (a). With the differentiable transformation, we are able to randomly generate two augmented images $t_1(\mathbf{X})$ and $t_2(\mathbf{X})$ that are respectively fed into the student and teacher networks f and $\hat{f}$. The two prediction posteriors can be then aligned via the differentiable inverse transformation $t_1^{-1}(f(t_1(\mathbf{X})))$ and $t_2^{-1}(\hat{f}(t_2(\mathbf{X})))$. For the rest of the paper, we denote prediction posteriors after alignment as $g_{i1} = t_1^{-1}(f(t_1(\mathbf{X}_i)))$ (student) and $\hat{g}_{j2} = t_2^{-1}(\hat{f}(t_2(\mathbf{X}_j)))$ (teacher).

D. Avoiding Collapsed Predictions on Unlabelled Data

Consistency-based SSL methods commonly use mean square error (MSE) for the consistency loss on unlabelled data. In this section, we first point out that MSE loss allows collapse of model predictions on unlabelled data to the majority class as



(a) (b)

Fig. 3. (a) Illustration of differentiable transformation. Red solid box and yellow dashed box indicate transformed and original views. The final yellow solid box indicates the prediction after differentiable inverse transformation where consistency loss is computed. (b) Loss surface for MSE consistency loss. When both student and teacher produce exactly the same result (may not be correct), as indicated by the diagonal, the MSE loss is always zero.

this is a trivial solution of the MSE consistency loss. We then introduce the N-pair loss as a way to mitigate this issue thus enabling the unlabelled data to better regularize the model.

Without loss of generality, suppose we have two scalar predictions (pixel-wise predictions) g_1 and g_2. We visualize the loss surface for MSE pairwise consistency in Fig. 3 (b), and observe that the MSE loss is flat along the diagonal because MSE loss is minimized when the predictions from the two networks match. When the class distribution is highly imbalanced as in the case of curvilinear segmentation, where there are far more background pixels than foreground pixels, the two networks can achieve zero MSE consistency loss just by assigning every single pixel to the majority class. In this case every pixel will be predicted as background. We term this all majority class prediction as a collapsed prediction. An instance of this behaviour is shown in Fig. 8 (a) where training IoU gets closer to 1, yet validation IoU collapses to 0. Avoiding this often requires very careful selection of the EMA hyperparameter and consistency weights, which requires expensive tuning runs.

In this work, inspired by the recent progress in contrastive learning [14], we propose to use an N-pair loss on unlabelled data [13] to avoid collapsed predictions. The N-pair loss simultaneously exploits all unlabelled samples in a training mini-batch to construct one positive pair to encourage similarity and multiple negative pairs to encourage diversity in predictions. By enforcing diversity with the negative pairs, the N-pair loss effectively penalizes models that give collapsed predictions on all unlabelled data. We illustrate the construction of the N-pair loss for curvilinear structure segmentation in Fig. 4: in a mini-batch of N_B unlabelled images the positive pair is chosen as the predictions from student and teacher on an anchor image $\mathbf{X}_i$, $\{(g_{i1}, \hat{g}_{i2})|i = 1 \cdots N_B\}$, while negative pairs are chosen as the predictions between the student's predictions on the same anchor image and teacher's predictions on other images in the mini-batch, $\{(g_{i1}, \hat{g}_{j2})|i \neq j; i, j = 1 \cdots N_B\}$. Formally, the N-pair loss is given by,

$$l_{\text{N-pair}} = -\frac{1}{N_B}\sum_{i=1}^{N_B}\log\frac{\exp(sim(g_{i1}, \hat{g}_{i2})/\tau)}{\sum_{j=1}^{N_B}\exp(sim(g_{i1}, \hat{g}_{j2})/\tau)} \tag{5}$$

where N_B, τ and sim are the mini-batch size, temperature parameter and similarity metric, respectively. We use cosine similarity defined in Eq. (6) for the similarity metric sim, where **vec** vectorizes a matrix into a vector and δ is a small value to avoid division by zero.

$$sim(g_1, g_2) = \frac{\mathbf{vec}(g_1 + \delta)^\top \mathbf{vec}(g_2 + \delta)}{||\mathbf{vec}(g_1 + \delta)||_2 \cdot ||\mathbf{vec}(g_2 + \delta)||_2} \tag{6}$$

We now formally show how N-pair loss prevents collapsed predictions that can occur when MSE is used. Specifically, we show that the N-pair loss will be lower for correct predictions than collapsed predictions, unlike MSE loss. First consider the case when predictions all collapse to a single class, often the background resulting in all zero predictions, i.e. $g_1 = \mathbf{0}$ and $\hat{g}_2 = \mathbf{0}$. Then, the cosine similarity for all positive and negative pairs are 1; the same result holds when predictions are all foreground, i.e. $g_1 = \mathbf{1}$ and $\hat{g}_2 = \mathbf{1}$. Then, the N-pair loss for this collapsed prediction is

$$\tilde{l}_{\text{N-pair}} = -\frac{1}{N_B} * N_B \log 1/N_B = \log N_B. \tag{7}$$

We note that in this case, the MSE loss is $\tilde{l}_{mse} = 0$. When the predictions are all correct, the positive pair similarity is 1, and for negative pairs this is approximately $\epsilon \ll 1$, because the segmentation masks for arbitrary two images are unlikely to significantly match, and the N-pair loss becomes

$$\begin{aligned} l_{\text{N-pair}} &= -\frac{1}{N_B} * N_B(\log\frac{\exp 1/\tau}{\exp 1/\tau + (N_B - 1) * \exp\epsilon/\tau}) \\ &= \log(1 + (N_B - 1) * \exp\frac{\epsilon - 1}{\tau}) \end{aligned} \tag{8}$$

while the MSE loss is still $l_{mse} = 0$. Since $\exp\frac{\epsilon-1}{\tau} < 1$, it is easy to verify that $l_{\text{N-pair}} < \tilde{l}_{\text{N-pair}}$ when N_B is larger than 1 and the ratio $l_{\text{N-pair}}/\tilde{l}_{\text{N-pair}}$ is getting smaller with increased batchsize N_B. This suggests the gap between N-pair loss under collapsed prediction and correct prediction is more significant with larger batchsize. We therefore conclude that the N-pair loss can easily distinguish good predictions from collapsed ones as opposed to MSE loss, for which the loss is 0 in both cases.

E. Modelling Spatial Correlation With Positional Encoding

The nature of curvilinear structure segmentation requires the network to encode spatial correlation. For example, cracks, roads, and blood vessels are spatially continuous thus pixels adjacent to other positive (foreground) pixels are likely to be

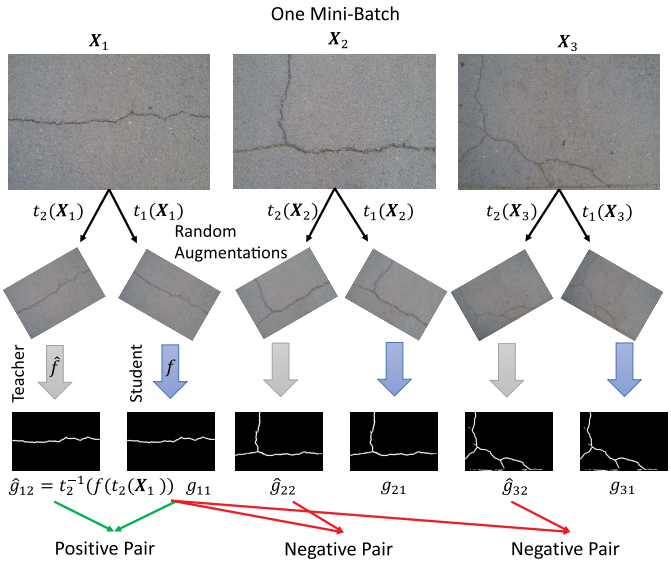


Fig. 4. Illustration of N-pair loss for unlabelled images. Positive pairs are constructed between the predictions of teacher and student networks on the same input image. Negative pairs are constructed between the predictions of teacher and student networks on different input images.

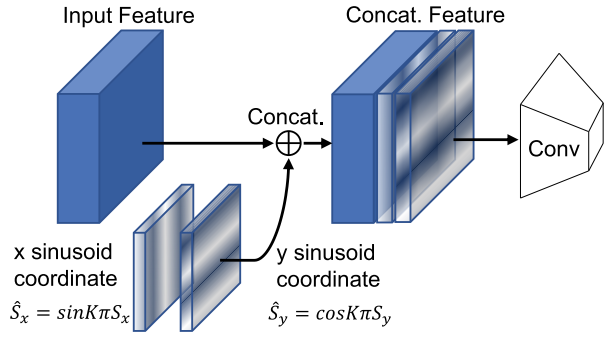


Fig. 5. Sinusoid coordinate encoding for capturing spatial correlation. Encodings are concatenated to feature maps in channel-wise fashion.

positive as well. Though it is difficult to handcraft features that capture this correlation, encoding spatial locations has been shown to be effective [16]. One straightforward way to encode spatial locations is by appending linear coordinates, normalized to between 0 and 1, to the intermediate feature layers as additional feature channels. However, learning directly on this absolute positional encoding risks overfitting to specific locations, especially when training with very few labelled samples. Inspired by the positional encoding of [35], we apply a sinusoid coordinate encoding as in Eq. (9) where S_x, S_y are the linear positional encoding normalized to between 0 and 1 and K is a period parameter. Such periodic positional encoding allows the network to be aware of relative location while reducing the risk of overfitting to absolute locations. The positional encodings are channel-wise concatenated to each intermediate activation map of the encoder network as shown in Fig. 5.

$$\hat{S}_x = \sin K\pi S_x, \quad \hat{S}_y = \cos K\pi S_y \tag{9}$$

Algorithm 1 Semi-Supervised Curvilinear Structure Segmentation: Training Algorithm

Input : Labelled Images $\{(\mathbf{X}_i^l, \mathbf{Y}_i^l)\}$ and Unlabelled Images $\{\mathbf{X}_j^u\}$
Output: Segmentation Model Parameters Θ
for $T \leftarrow 1$ **to** 1000 **do**
  for $\{\mathbf{X}_i^l, \mathbf{Y}_i^l, \mathbf{X}_j^u\}$ *in mini-batch* **do**
    // Random Augmentation
    Compute $t(\mathbf{X}_i^l), t_1(\mathbf{X}_i^u), t_2(\mathbf{X}_i^u)$;
    // Forward Pass
    Compute $f(t_1(\mathbf{X}_i^l)), f(t_1(\mathbf{X}_i^u)), \hat{f}(t_2(\mathbf{X}_i^u))$
  // Supervised Loss
  Compute l_{dice} according to Eq. (10);
  // Unsupervised Loss
  Compute $l_{\text{N-pair}}$ according to Eq. (5);
  // Unlabelled Loss Weight
  $\gamma = \exp(-10(\frac{min(T,T_{rp})}{T_{rp}} - 1)^2)$;
  // Train One Minibatch
  $\Theta = \Theta - \alpha \nabla_\Theta l_{total}$;

F. Dice Loss for Class Imbalance

It is well-known that class imbalance can affect the performance of machine learning models [36]–[38]. The cross-entropy (CE) loss commonly used for supervised segmentation tasks averages the pixel-wise CE loss over all pixels in the image; CE loss biases learning towards the majority class when there is an imbalanced class distribution. In curvilinear structure segmentation tasks, this imbalance is particularly severe due to a low positive class ratio (see Table I). As a result, models trained with CE may have high per-pixel accuracy but low intersection over union (IoU), which is the most common evaluation metric for segmentation tasks. In order to mitigate this issue, Dice loss [39] (Eq. 10) was proposed as an alternative for segmentation; it focuses on the intersection of predictions with ground truth only for the positive class.

$$l_{dice} = 1 - \frac{2\mathbf{vec}(g_1 + \delta)^\top \mathbf{vec}(\mathbf{Y} + \delta)}{||\mathbf{vec}(g_1 + \delta)||_1 + ||\mathbf{vec}(\mathbf{Y} + \delta)||_1} \tag{10}$$

G. Implementation Details

The final loss function is the weighted combination of supervised loss and unlabelled loss: $l_{total} = l_{dice} + \gamma\, l_{\text{N-pair}}$. As commonly done in consistency-based SSL, we do not apply the unlabelled loss $l_{\text{N-pair}}$ at the beginning of training but instead gradually increase the strength of γ following a sigmoid function $\gamma = \exp(-10(\frac{min(T,T_{rp})}{T_{rp}} - 1)^2)$ where T and T_{rp} are epoch number and hyperparameter, respectively. We set the mean teacher EMA hyperparameter α to 0.999. For all experiments, we use the SGD optimizer, fixing the total number of training epochs to 1000 and $T_{rp} = 500$, the initial learning rate at 0.001 and continuously decay by half every 500 epochs. Each epoch is defined as cycling all labeled data once. The sinusoid spatial encoding parameter was set as $K = 4$. The overall algorithm is summarized in Algorithm 1.

TABLE I
PERCENTAGE (%) OF POSITIVE PIXELS IN DATASETS

Dataset	CrackForest	Crack500	Gaps384	MITRoad	EM128	DRIVE128
Pos. pixel %	2.5%	6.0%	1.2%	5.1%	22.0%	8.6%

IV. EXPERIMENTS

We validate the efficacy of SemiCurv on 6 different curvilinear structure datasets spanning 3 domains: road crack segmentation, road segmentation from satellite images, and biomedical image segmentation.

A. Datasets

CrackForest [40] was proposed for crack segmentation from paved road images. It consists of 118 labelled images in total. We created a datasplit for evaluating semi-supervised segmentation by randomly splitting the whole dataset into 80% for training, 10% for validation, and 10% for testing. Among the training images, we further assume 5% or 1% of the images are labelled. **Crack500** [1] is a more comprehensive dataset consisting of 1896/348/1124 labelled road crack images for training, validation, and testing respectively. We follow the standard datasplit proposed in [1] and assume the training set is partially labelled. **Gaps384** [1] was created from a large road crack detection dataset GAPS [41] by manually selecting 384 images and cropping out smaller regions with cracks, totalling 508 annotated images. We follow the standard data split for evaluation. **MIT Road** [18] was proposed for automatically extracting road from satellite images. As we notice there are many blank regions in the RGB images, we preprocess the images to manually crop non-blank regions. In total, we obtain 6880/1215/440 labelled images for training, validation, and testing respectively. **EM128** [23] was created for evaluating segmentation on cell membranes. It consists of 30 labelled images and we follow the practice in [42] to reserve 15 images for training and the rest for testing. To allow evaluation of semi-supervised learning, we divide each image into 16 regions each covering 128×128 pixels. In total we have 240/240 for training and testing respectively and we name the derived dataset as EM128. **DRIVE128** The DRIVE dataset [43] was developed for evaluating segmentation on retinal blood vessels. It consists of 20 labelled images which we split into 10 for training and 10 for testing. Non-overlapping patches of size 128×128 pixels are cropped in a similar way to EM128 and we denote this derived dataset as DRIVE128. The overall number of training, validation, and test samples are shown in Table II. The number of labelled training data at different levels of supervision are provided as well. We note that the level of class imbalance is severe on most of these datasets, as can be seen from the low percentage of positive (foreground) pixels given in Table I. The high class imbalance potentially makes the model more likely to collapse to predicting all pixels as background on the unlabelled data.

B. Evaluation Metric

To evaluate the quality of segmentation predictions, we compute Intersection over Union (IoU) using a threshold of 0.5 on prediction posteriors for all test images and report the mean IoU. We also evaluate the F1 measure of precision and recall at threshold 0.5. This metric treats segmentation as a binary classification task. For all 6 datasets, we set the background pixel label as 0 and foreground as 1.

C. Competing Methods

We extensively compare SemiCurv with existing fully supervised methods (with different backbone networks) and state-of-the-art generic semi-supervised learning methods, as described in the following.

1) Fully Supervised Methods: We first investigate the state-of-the-art fully supervised curvilinear structure segmentation, edge detection, and semantic segmentation methods. This provides context of what is achievable for curvilinear segmentation tasks. **FPHBN** [1] is the state-of-the-art method for crack segmentation; we used the reported results on Crack500 in this paper for comparison. **HED** [44] was proposed for detecting edges in images. Due to the similar nature of edge and linear structures defined in this work, we evaluate this backbone as a baseline. **DeepLab v3+** [45] is the state-of-the-art backbone network for semantic segmentation tasks. We evaluate this network here to show that generic semantic segmentation networks do not generalize well to curvilinear structure segmentation. **UNet** [21] was originally proposed for medical image segmentation. We use a variant that adds residual connections in each convolution block as the backbone network in this work. Details of the backbone network is given in the Supplementary Material.

2) Semi-Supervised Methods: We compare against state-of-the-art semi-supervised semantic segmentation, medical image segmentation, and generic SSL methods adapted to curvilinear segmentation. **CutMix** [9] proposed to generate random mask to mix two images and consistency is applied to between the predictions of teacher model and student model over the masked regions. **TCSM** [33] adapts the mean teacher framework to learn from unlabelled data. The augmentation is limited to scaling and rotation in multiples of 90°. **VAT** [27] proposed to learn the optimal augmentation by maximizing the consistency on unlabelled data. The final objective aims to minimize the pairwise consistency as well as the entropy on unlabelled data. **cGAN** [46] is another line of semi-supervised semantic segmentation approach. The conditional GAN (cGAN) based method introduced a discriminator network to differentiate ground-truth segmentation mask from predicted ones and the predictions on unlabeled data can be constrained by the discriminator. **Mean Teacher (Baseline)** [26] proposed to use a temporal ensemble, as teacher, of a learnable student model to provide pseudo labels to train the student model. MSE consistency is used between teacher and student outputs. We use this as our baseline method. **SemiCurv** is our proposed model incorporating differentiable geometric transformations, N-pair objective, and sinusoid positional encoding.

D. Quantitative Results

We present the comparison of both fully supervised and semi-supervised methods with 5% and 1% labelled data

TABLE II
SEMI-SUPERVISED LEARNING DATA SPLITS FOR ALL 6 DATASETS

Split		CrackForest		Crack500		Gaps384		DRIVE128		MITRoad		EM128	
Train	Labelled Percent.	5%	1%	5%	1%	5%	1%	5%	1%	1%	0.1%	10%	5%
	#Labeled	5	1	95	19	23	5	8	2	69	7	24	12
	#Unlabeled	91	95	1801	1877	442	460	152	158	6811	6873	216	228
Test		11		1124		39		160		440		240	

TABLE III
QUANTITATIVE EVALUATIONS OF SEMI-SUPERVISED CURVILINEAR SEGMENTATION. ALL NUMBERS ARE IN %. (*95/90%) INDICATES MULTIPLIED BY 95/90%

		CrackForest		Crack500		Gaps384		DRIVE128		MIT Road		EM128	
	Model	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Fully Supervised	Labelled Data	100%		100%		100%		100%		100%		100%	
	DeepLabV3	51.2	67.1	38.8	52.5	37.3	53.1	30.5	33.6	33.2	48.3	46.0	61.8
	HED	54.7	70.2	39.5	53.9	35.2	49.0	33.6	37.6	40.1	55.0	46.6	63.0
	FPHBN	NA	NA	48.9	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Unet 100%	70.7	84.7	49.6	62.9	39.0	52.9	59.2	60.4	58.4	71.3	67.5	80.2
	Unet 95%	67.1	80.4	47.1	59.7	37.1	50.3	56.2	57.3	55.5	67.7	64.1	76.2
	Unet 90%	63.6	76.4	44.6	56.6	35.1	47.8	53.3	54.5	52.6	64.3	60.7	72.4
Semi-Supervised	Labelled Data	5%		5%		5%		5%		1%		10%	
	Unet	60.1	74.3	46.5	60.5	34.9	48.9	49.1	56.5	53.8	67.3	63.4	77.1
	CutMix	66.0	79.3	44.4	57.5	33.5	46.9	50.3	49.5	55.8	69.0	62.2	76.3
	VAT	53.9	69.1	48.1	62.5	29.6	42.3	47.8	56.3	48.3	62.6	57.9	72.6
	TCSM	49.9	66.2	36.8	50.7	37.8	52.2	44.8	51.0	32.5	47.0	33.0	48.8
	MT	66.9	79.1	48.0	61.9	34.3	48.2	54.2	60.9	55.2	68.7	65.3	78.6
	cGAN	67.0	79.2	48.1	62.0	33.7	47.3	55.5	62.4	53.5	66.6	62.1	74.8
	Ours	**68.0**	**80.9**	**49.2**	**62.4**	**38.0**	**53.4**	**56.0**	**61.3**	**56.0**	**69.1**	**65.7**	**79.1**
	Labelled Data	1%		1%		1%		1%		0.1%		5%	
	Unet	36.6	48.8	44.4	58.3	28.3	41.2	34.1	45.5	35.9	50.5	57.0	71.5
	CutMix	56.5	71.2	40.2	54.6	30.1	43.6	43.0	50.8	44.9	59.2	63.9	77.6
	VAT	29.3	41.5	43.3	58.4	18.8	17.3	31.4	41.9	25.4	38.6	57.1	72.2
	TCSM	30.5	43.9	41.9	56.2	28.6	41.7	28.8	34.6	31.6	45.4	28.6	43.6
	MT	61.9	75.9	45.6	59.8	31.1	44.3	37.7	47.3	46.4	61.1	63.2	76.9
	cGAN	54.6	66.9	45.4	59.5	30.1	42.8	42.6	53.4	43.6	57.4	60.3	73.4
	Ours	**64.4**	**78.1**	**46.2**	**60.1**	**33.7**	**48.4**	**44.3**	**51.5**	**47.7**	**62.2**	**64.1**	**77.9**

in Table III; for MITRoad and EM128 we consider different labeling budgets due to the size of dataset. In the fully supervised block, *UNet* (∗95%) and *UNet* (∗90%) indicate the relative performance $95\% * m$ and $90\% * m$ where m is the performance achieved by the fully supervised method trained with 100% of the labelled data. From the extensive comparison, we make the following observations. First, the adapted *UNet* is a very strong backbone network for curvilinear structure segmentation. It outperforms the state-of-the-art semantic segmentation backbone *DeepLabV3+*, edge detection network *HED*, and network specifically designed for crack segmentation *FPHBN* on all datasets. Moreover, under the semi-supervised setting, with only 5% and 1% labelled data *SemiCurv* can match and even outperform the 95% and 90% relative performance of fully supervised counterparts in terms of IoU and F1, respectively. By comparing to the state-of-the-art semi-supervised methods for segmentation, we still observe very significant advantages for *SemiCurv*. In particular, the improvement from baseline is more significant in the lower label regime. For example, on CrackForest IoU improved 17% and 2.2% from the *UNet* baseline trained using 1% and 5% labelled data respectively. The improvement on all other datasets also exhibits similar patterns with large improvements over the UNet baseline. We also observe *CutMix* to be a very competitive method, in particular on the MIT Road dataset. We speculate that this is because the MIT Road dataset resembles generic semantic segmentation problems, such that it benefits from cutmix style augmentation. Finally, *TCSM* produces much worse results compared to both *MT* and *SemiCurv*. We speculate that this is due to weak augmentation adopted in *TCSM* hampering semi-supervised performance.

TABLE IV
COMPARISON OF POSITION EMBEDDINGS

CrackForest	Sinusoid Enc. (ours)	Learned [35]	Rotary [47]
1%	64.41	62.93	62.96
5%	68.02	67.49	67.56

E. Ablation Study

Here we carry out ablation studies to investigate the effectiveness of each component and present the results in Table V.

TABLE V
ABLATION STUDY FOR ALL 6 CURVILINEAR DATASETS

SSL	SupLoss	GeoTform	Consist.	Spt.Enc.	Sim.	CrackForest 1% IoU	Crack500 1% IoU	Gaps384 1% IoU	MIT Road 1% IoU	EM128 5% IoU	DRIVE128 5% IoU
-	WBCE	✓	-	-	-	21.1	38.6	24.3	41.7	40.9	34.9
-	Dice	✓	-	-	-	36.6	44.4	28.3	51.6	57.0	49.1
MT	Dice	-	MSE	-	-	50.5	43.2	19.2	51.4	53.8	43.9
MT	Dice	✓	MSE	-	-	61.9	45.6	31.1	55.2	63.2	54.2
MT	Dice	✓	N-pair	-	Cosine	63.7	45.9	31.6	55.2	63.9	55.9
MT	Dice	✓	N-pair	Linear	Cosine	62.8	45.9	**34.0**	54.5	63.7	44.9
MT	Dice	✓	N-pair	Sinusoid	Cosine	**64.4**	**46.2**	33.7	**56.0**	**64.1**	**56.0**
MT	Dice	✓	N-pair	Sinusoid	L2	58.4	45.2	29.5	53.0	55.9	43.0

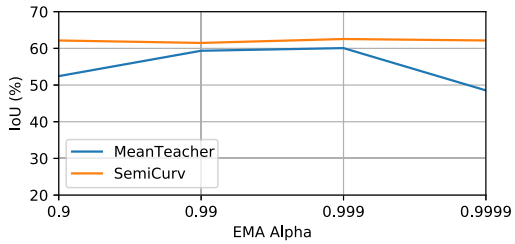


Fig. 6. Evaluation of EMA hyperparameter.

1) Supervised Loss: We first compare the supervised loss, *SupLoss*, adopted for labelled images. *UNet* with weighted binary cross entropy loss, *WBCE*, is consistently worse than *UNet* with dice loss, *Dice*. This suggests that optimizing dice loss is able to better generalize under severe class imbalance.

2) Semi-Supervised Learning: We next compare the fully supervised baseline with the mean teacher model, *MT*. With mean square error (MSE) consistency loss, we observe significant improvement of *MT* over the baseline. We further evaluate *MT* with and without the differentiable geometric transformation, *GeoTform*. The significant difference in performance between the two settings indicates that strong augmentation is vital to the success of consistency-based SSL.

3) MSE vs N-Pair Consistency Loss: We further compare *MT* with mean square error loss, *MSE*, and N-pair loss. The improvement demonstrates the superiority of considering both positive and negative pairs simultaneously with *N-pair*. In addition to the numerical improvement, we also show in Fig. 8 (a) that *N-pair* is more robust than *MSE* and avoids collapsed predictions. As we discussed in Sect. III-D, the pairwise mean square error, *MSE*, consistency loss is prone to a collapsing issue. This becomes severe when the data is highly imbalanced and very few labelled samples are available for training. We present examples for both *MSE* consistency and *N-pair* consistency on CrackForest with only 1% labelled data in Fig. 8 (a). *MSE* produces unstable validation performance, and it eventually collapses to all zeros around 800 epochs. In contrast, with the proposed *N-pair* loss the validation performance keeps stable throughout the training. Overfitting, due to too few labelled data, is still observed but it avoids collapsing to all zero predictions.

4) Positional Encodings: We analyze the effect of including various forms of spatial encoding. First, when including linear spatial encoding, *Linear Spt. Enc.*, we sometimes observe that performance degrades. This can be attributed to the potential overfitting to absolute locations; using sinusoid spatial encoding instead, *Sinusoid Spt. Enc.*, we observe clear and consistent improvements on all datasets. We investigate the impact of constant K in Eq. 9 which controls the period of sinusoid positional encoding. The final *SemiCurv* is evaluated with $K = 1, 2, 4, 8$ with results shown in Fig. 8 (b). We observe an optimal range of K around 4, and overall the results are robust to K from 2 to 8. This shows sinusoid positional encoding is a robust component of our approach.

For alternative positional encodings, we evaluate two additional methods on CrackForest segmentation task. The results are presented in Tab. IV. We observe a small drop of performance when swapping out our proposed sinusoid encoding with learned position embedding [35] and rotary position embedding [47]. Overall, the proposed sinusoid position encoding is still the better option.

5) Alternative Similarity Metric: We further explore a different option for the similarity metric used in the N-pair loss. L2 distance is often used in N-pair loss for metric learning and we therefore evaluated the induced RBF kernel for comparison against cosine similarity. Formally, the L2 distance induced similarity is given by

$$l_{\text{N-pair}} = -\frac{1}{N_B}\sum_{i=1}^{N_B}\log\frac{\exp(-||g_{i1}-\hat{g}_{i2}||_2)}{\sum_{j=1}^{N_B}\exp(-||g_{i1}-\hat{g}_{j2}||_2)} \quad (11)$$

The quantitative comparison between using cosine similarity and L2 distance induced similarity is shown in Table V under the similarity metric (Sim.) column. For both similarity metrics, we use the SemiCurv framework and keep the training protocols unchanged. We observe from the comparison that cosine similarity outperforms L2 distance consistently on all datasets. Also, stability of training suffers with L2 distance compared to SemiCurv with cosine distance. This can be attributed to the normalizing effect of cosine similarity, while L2 distance is affected by the absolute number of positive pixel predictions.

6) Time Complexity: We analyze the time complexity of proposed semi-supervised learning method. With more unlabeled data, it takes more time to allow the network to iterate

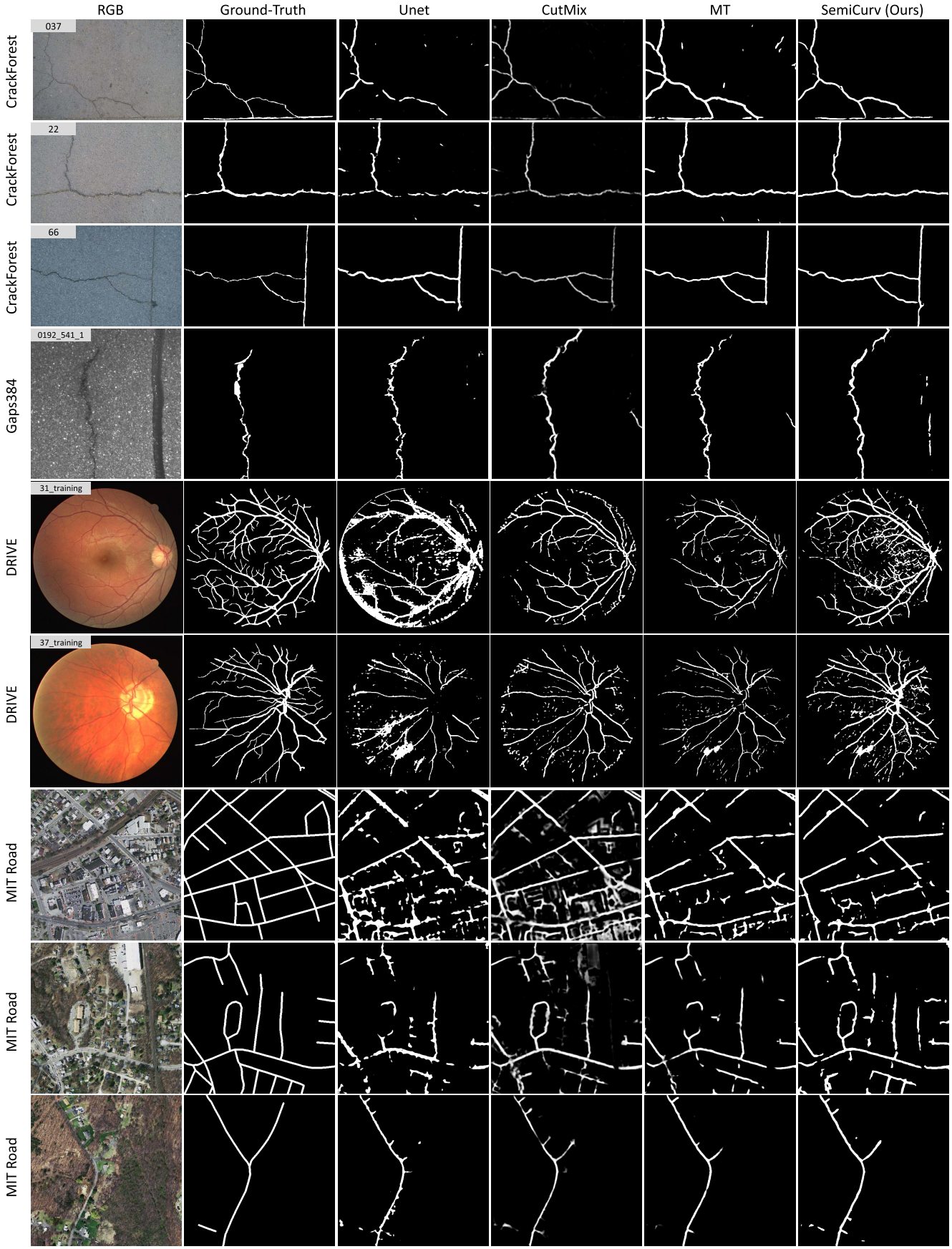
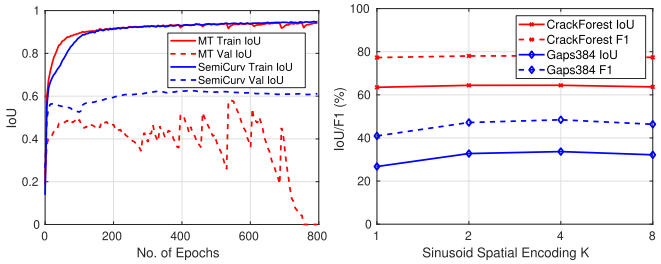


Fig. 7. Qualitative comparisons for curvilinear structure segmentation.



(a) (b)

Fig. 8. (a) Comparing MSE loss v.s. N-pair loss over collapsed predictions. (b) Performance v.s. sinusoid parameter K.

TABLE VI
EMPIRICAL EVALUATION OF TIME COMPLEXITY FOR PROPOSED METHOD

CrackForest	FullSup	SemiCurv
1%	20m	1h21m
5%	100m	5h30m

unlabeled data. The training time is roughly linearly proportional to the ratio of unlabeled and labeled data. We further empirically evaluated the training time for fully supervised learning method and SemiCurv in Tab. VI. Despite more training time required, our SemiCurv performs substantially better at inference stage and there is no additional cost at inference stage.

7) Exponential Moving Averaging Hyperparameter: We choose the EMA hyperparameter for teacher model according to the validation performance. A lower α will allow faster update of teacher network while a higher α slows down the update of teacher network. This hyperparameter is particularly sensitive for MSE consistency loss according because MSE loss is more likely to produce collapsed prediction, i.e. predicting all pixels as background. In this section we further provide empirical evidence for the chosen hyperparameter α by varying α from 0.9 to 0.9999 and compare the performance on CrackForest dataset. The results in Fig. 6 suggest choosing $\alpha = 0.999$ is the optimal choice for both mean teacher and our SemiCurv models.

F. Qualitative Results

We further present qualitative comparisons of *UNet*, *CutMix*, *MT*, and *SemiCurv* on 4 datasets in Fig. 7. We make the following observations: first, for both *MT* and *SemiCurv*, the visual quality of segmentation results are more appealing and better mimicks the ground-truth compared to the fully supervised baseline *UNet*. In particular, our method is able to generate more well-connected predictions for 037 (1st row) thanks to the contribution of appending positional encoding. We further notice that *SemiCurv* produces, in general, cleaner outputs for 037 (1st row), 022 (2nd row), 066_1 (3rd row). We also observe, in 0192_541_1 (4th row), the proposed approach captures very subtle cracks on the top. The segmentation results on DRIVE dataset (5th/6th rows) also suggest the superiority of *SemiCurv*. It produces relatively high fidelity and well-delineated blood vessel segmentation maps. In comparison, *UNet*, *CutMix*, and *MT* tend to mingle different blood vessels together which is less common in *SemiCurv*'s predictions. For satellite image segmentation, the visualization covers industrial areas (row 7) and rural areas (rows 8-9). We observe consistent improvements of *SemiCurv* compared to both *UNet* and state-of-the-art semi-supervised image segmentation *CutMix*. Interestingly, the *SemiCurv* predictions sometimes contain valid road branches that are missing in the ground-truth annotation (last row).

V. CONCLUSION

In this work we addressed curvilinear structure segmentation in a semi-supervised learning setting to exploit "freely" available and abundant amounts of unlabelled data with our proposed SemiCurv framework. In particular, we introduced stronger augmentation involving affine geometric transformations which we showed to be essential to the success of SSL. We further identified implications of the severe class imbalance in curvilinear segmentation tasks on the widely used MSE consistency loss, and showed that N-pair loss should be used instead to mitigate these issues. Finally, we found that sinusoid positional encoding is effective in further improving segmentation performance. Our extensive experiments on 6 curvilinear structure segmentation datasets from 3 different domains demonstrates the effectiveness of SemiCurv, and detailed ablation studies validated the importance of each of our proposed components. Our SemiCurv framework and the extensive results presented in this work provide a foundation for future work in SSL for curvilinear segmentation.

REFERENCES

[1] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.

[2] F. Wang, Y. Gu, W. Liu, Y. Yu, S. He, and J. Pan, "Context-aware spatio-recurrent curvilinear structure segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12648–12657.

[3] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-preserving deep image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5657–5668.

[4] A. Mosinska, P. Marquez-Neila, M. Kozinski, and P. Fua, "Beyond the pixel-wise loss for topology-aware delineation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3136–3145.

[5] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[6] W. C. Hung, Y. H. Tsai, Y. T. Liou, Y. Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[7] A. K. Mondal, A. Agarwal, J. Dolz, and C. Desrosiers, "Revisiting CycleGAN for semi-supervised segmentation," 2019, *arXiv:1908.11569*.

[8] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12671–12681.

[9] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," in *Proc. Brit. Mach. Vis. Conf.*, 2020.

[10] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1369–1378.

[11] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[13] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[16] R. Liu *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9628–9639.

[17] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.

[18] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[19] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in SAR satellite images with deep fully convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1867–1871, Dec. 2018.

[20] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[22] F. Isensee *et al.*, "NNU-Net: Self-adapting framework for u-net-based medical image segmentation," in *Bildverarbeitung Für die Medizin*. Cham, Switzerland: Springer, 2019.

[23] I. Arganda-Carreras *et al.*, "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Frontiers Neuroanatomy*, vol. 9, p. 142, 2015.

[24] 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.

[25] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[27] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[28] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3546–3554.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[30] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–15.

[31] A. Odena, "Semi-supervised learning with generative adversarial networks," 2016, *arXiv:1606.01583*.

[32] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5688–5696.

[33] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[34] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6256–6268.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[36] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Proc. 4th Int. Conf. Natural Comput.*, Oct. 2008, pp. 192–201.

[37] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[38] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[39] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017.

[40] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.

[41] M. Eisenbach *et al.*, "How to get pavement distress detection ready for deep learning? A systematic approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2039–2047.

[42] M. Seyedhosseini, M. Sajjadi, and T. Tasdizen, "Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2168–2175.

[43] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[44] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[45] 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.*, Sep. 2018, pp. 801–818.

[46] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[47] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," 2021, *arXiv:2104.09864*.