An Iterative Semi-supervised Approach with Pixel-wise Contrastive Loss for Road Extraction in Aerial Images

HUIJIE ZHANG, San Diego State University, USA and University of California, Santa Barbara, USA
PU LI and XIAOBAI LIU, San Diego State University, USA
XIANFENG YANG, Civil and Environmental Engineering, University of Maryland, USA
LI AN, San Diego State University, USA

Extracting roads in aerial images has numerous applications in artificial intelligence and multimedia computing, including traffic pattern analysis and parking space planning. Learning deep neural networks, though very successful, demand vast amounts of high-quality annotations, of which acquisition is time-consuming and expensive. In this work, we propose a semi-supervised approach for image-based road extraction in which only a small set of labeled images are available for training to address this challenge. We design a pixel-wise contrastive loss to self-supervise the network training to utilize the large corpus of unlabeled images. The key idea is to identify pairs of overlapping image regions (positive) or non-overlapping image regions (negative) and encourage the network to make similar outputs for positive pairs or dissimilar outputs for negative pairs. We also develop a negative sampling strategy to filter false-negative samples during the process. An iterative procedure is introduced to apply the network over raw images to generate pseudo-labels, filter and select high-quality labels with the proposed contrastive loss, and retrain the network with the enlarged training dataset. We repeat these iterative steps until convergence. We validate the effectiveness of the proposed methods by performing extensive experiments on the public SpaceNet3 and DeepGlobe Road datasets. Results show that our proposed method achieves state-of-the-art results on public image segmentation benchmarks and significantly outperforms other semi-supervised methods.

CCS Concepts: • **Computing methodologies** → **Image segmentation**; • **Theory of computation** → *Semi-supervised learning;*

Additional Key Words and Phrases: Deep learning, semi-supervised learning, contrastive loss, iterative labeling, road extraction

ACM Reference format:
Huijie Zhang, Pu Li, Xiaobai Liu, Xianfeng Yang, and Li An. 2023. An Iterative Semi-supervised Approach with Pixel-wise Contrastive Loss for Road Extraction in Aerial Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 3, Article 80 (November 2023), 21 pages.
https://doi.org/10.1145/3606374

This work was supported by the National Science Foundation (grant no. 2106965).
Authors' addresses: H. Zhang, San Diego State University, 5500 Campanile Dr, San Diego, California, USA and University of California, Santa Barbara, California, USA; email: hzhang9970@sdsu.edu; P. Li, X. Liu, and L. An, San Diego State University, 5500 Campanile Dr, San Diego, California, USA; emails: {pli5270, xiaobai.liu, lan}@sdsu.edu; X. Yang, Civil and Environmental Engineering, University of Maryland, College Park, MD, USA; email: xtyang@umd.edu.

1551-6857/2023/11-ART80 $15.00
https://doi.org/10.1145/3606374

1 INTRODUCTION

With the rapid development of remote sensing technology, high-quality, fine-resolution aerial and satellite images can be acquired at much lower cost and provide promising resources for various applications, including emergency management, urban planning, and disaster monitoring. Road extraction is one of the most critical tasks that utilize these image data and plays a vital role in autonomous driving, road condition evaluation, Global Positioning System (GPS) navigation, and intelligent city management [1, 30]. In the literature, deep learning–based approaches have shown great potential in road extraction from aerial and satellite images [5, 6, 23, 24, 38, 46, 55]. However, most of these methods are based on supervised learning, which highly relies on the quantity and quality of annotations. More data help improve the precision and recall of a supervised model for road extraction. Nevertheless, creating and maintaining images with pixel-wise road labels is time-consuming and expensive. In this work, we will develop a learning-based method to leverage large-scale raw satellite images while minimizing human efforts. Figure 1 shows the results of the proposed method over two satellite images.

Our work is relevant to past efforts on semi-supervised learning and its applications over semantic segmentation. These methods can maximize the use of unlabeled images and have gained encouraging successes on various image-based tasks, including image classification [2, 8, 57, 66], text classification [37, 65], object detection [31, 52, 64], and dehaze [35, 54, 60]. However, few studies employ semi-supervised methods for image-based road extraction [38]. This article introduces a semi-supervised iterative approach for road extraction, in which a small set of labeled images and a large number of unlabeled images are incorporated for training. As shown in Figure 2, our empirical study suggested that training a supervised road extraction model with fewer satellite images will lead to inferior performance. We propose an iterative learning framework to fully take advantage of the unlabeled image corpus. Our framework includes three iterative steps: initially training a semi-supervised model with a pixel-wise contrastive loss; applying the current network to the raw images to obtain their pseudo labels, which are used to enlarge the training set; and retraining a semi-supervised model with pseudo labels. We repeat the last two steps until convergence.

Our work is also inspired by the so-called contrastive learning, which has been proven to be a practical self-supervised approach for generic image tasks [11, 22]. It aims to learn a deep representation so that positive pairs of training images have similar representations, whereas the representations of negative pairs are dissimilar. Previous contrastive learning methods usually work on image-level features and minimize the distance between augmented views of the same image while maximizing the distance between different images [11, 22]. The same idea is not valid for pixel-wise image tasks, including image-based road extraction, because of the challenges of finding positive and negative pairs without access to pixel-level labels. In this work, we propose a simple yet effective way to collect positive pairs of image regions for roadway extraction. We first randomly crop an image region (e.g., 256 by 256 pixels) and expand this region into the adjacent areas to obtain a pair of overlapping image regions. These two enlarged regions are then fed into the network, and the network outputs are expected to be the same for the common area. This leads to the positive pairs of image regions. This method extracts the feature representations of the common area with different surrounding contexts and tries to minimize the divergence between these representations. Similarly, we will crop negative image regions that include different shapes of infrastructures and backgrounds. A contrastive loss function is then defined over both the positive and negative pairs to guide the training of the network. This pixel-wise contrastive learning scheme can be applied to both labeled and unlabeled images.

We further introduce an iterative scheme to take advantage of the raw images in the semi-supervised setting. Similar approaches were previously developed to apply self-supervised techniques to semi-supervised segmentation tasks [8, 10, 19, 26]. These methods usually first obtain

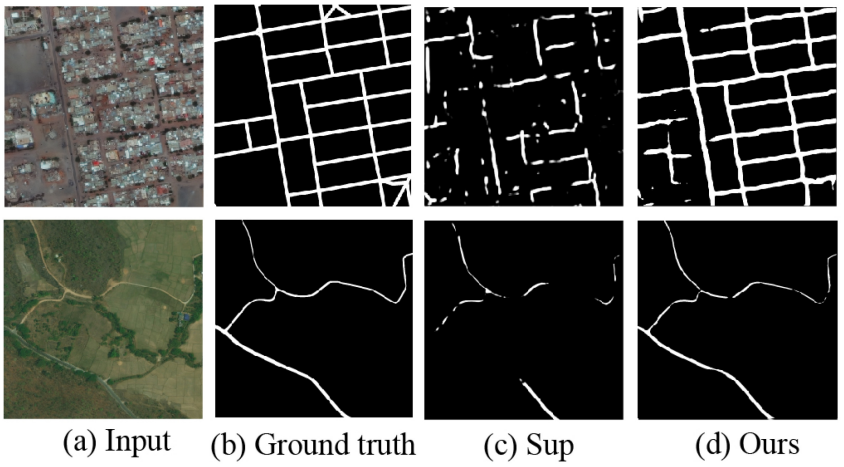

Fig. 1. Image-based road extraction. (a) Two aerial images; (b) ground-truth pixel-wise labels; (c) labels estimated by a supervised method [49]; (d) labels estimated by the proposed semi-supervised method. Both methods (c) and (d) have access to the same labeled data. Our method (d) utilizes extra unlabeled images.

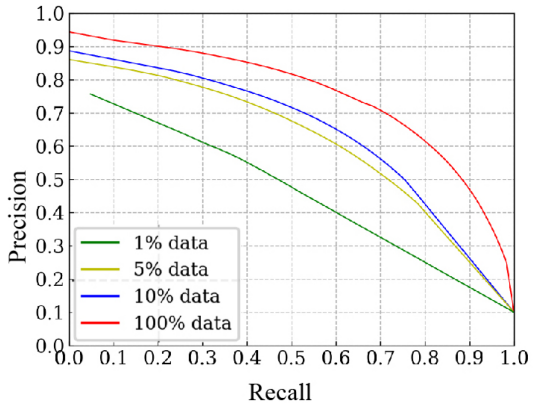

Fig. 2. Precision-Recall curve for U-Net [49] trained with different proportions of the labeled images on the dataset SpaceNet3 [56].

pseudo annotations from the model trained with labeled data and retrain the model repeatedly. One might also use curriculum learning [7] to select the most appropriate training samples at each training iteration. In this work, we develop an iterative labeling method that employs the proposed contrastive loss to rank raw images with pseudo labels and add the highly ranked pseudo labels into the labeled set. The iterative scheme leads to a progressive growing algorithm that can grow the labeled set with high-confident pseudo labels.

We conduct extensive experiments on the public SpaceNet3 [56] and DeepGlobe Road [17] datasets. Results suggest that the proposed semi-supervised method outperforms other semi-supervised methods. The following are the main technical contributions of this study.

- We develop an effective contrastive learning approach for extracting roadways in images. A pixel-wise contrastive loss is introduced to explore the contextual representation of local image regions and leverage both labeled and unlabeled training images.

- We propose an iterative self-training framework that can employ the proposed contrastive loss to identify high-quality pseudo labels and gradually grow the training dataset to boost model performance.
- The proposed method achieves state-of-the-art results on the public SpaceNet3 and DeepGlobe Road datasets and significantly outperforms the alternative supervised methods using the same amount of labeled images.

The rest of this articlde is organized as follows. Section 2 briefly reviews the related work. In Section 3, the proposed iterative semi-supervised approach with the pixel-wise contrastive loss for road extraction is illustrated in detail. Experimental results are presented in Section 4. Finally, we draw our conclusions in Section 5.

2 RELATIONSHIPS TO PREVIOUS WORK

This work is closely relevant to four research streams in the literature: image-based road extraction, semi-supervised semantic segmentation, contrastive loss, and self-supervised learning.

2.1 Image-Based Road Extraction

Numerous approaches have been explored to extract roads from aerial and satellite images. In the past few decades, traditional methods employed handcrafted image features to represent individual pixels [41, 48] or objects [42, 44] in images. Although traditional road extraction methods have achieved some good results, the robustness and generalization ability is limited as the features are manually designed and the procedure might not be generalized well to new testing images.

With the successes of deep neural networks, state-of-the-art methods usually cast road extraction as a segmentation problem and employ an end-to-end trained network to output pixel-wise labels directly. U-Net [49] introduced a downsampling path to obtain semantic information and a symmetric upsampling path to acquire localization information, and has been widely used for image-based road extraction [40, 71]. To further boost road extraction performance, cascaded neural networks are utilized to carry out a multistage semantic segmentation framework [14]. A Generative Adversarial Network (GAN) was also incorporated with road detection models [70]. Some recent works focus on improving connections of road networks [6, 43]. Batra et al. [6] proposed a supervised learning framework in which a multi-branch Convolutional Neural Network (CNN) is utilized to simultaneously learn road orientation and segmentation, in which the orientation branch aims to improve road connectivity.

Deep learning methods can be further integrated with graph-based representations of roadways to improve system accuracy. A classical way is to consider each road segment as a graph node and connect nodes following the road topology [5, 55]. The classical work RoadTracer [5] trains a CNN and adopts an iterative search process to capture connected road pixels to get the road graph.

Both segmentation-based and graph-based methods have achieved impressive performance for road extraction tasks. However, most of these methods use supervised learning techniques and require a large amount of annotated training data. A few studies employed semi-supervised road extraction methods [38]. Liu et al. [38] proposed a semi-supervised framework to extract roads, incorporating high-level feature selection, Markov Random Field (MRF), and ridge transversal method. Unlike the mentioned methods, our study focuses on the segmentation-based method to explore a practical semi-supervised approach with pixel-wise contrastive loss and iterative labeling for road extraction.

2.2 Semi-supervised Semantic Segmentation

Semi-supervised learning aims to employ a small set of labeled and a significant number of unlabeled data to enhance representation learning. In the literature, multiple efforts exist to train deep

networks in the semi-supervised setting [10, 13, 15, 26, 39, 59, 62]. One category of these methods added the so-called consistency regularization to the loss function to promote the consistency of predictions under different perturbations [20, 21, 33, 47]. One popular choice is to augment data and impose consistency regularization between the augmented and original samples. In particular, Yun et al. [20] applied CutMix over training images and imposed the consistency between the network outputs over mixed images and raw images. Ouali et al. [47] further presented a cross-consistency training (CCT) scheme, which enforces consistency between the outputs of the primary and affiliated decoder.

Generative models are also leveraged with consistency regularization for semi-supervised semantics segmentation [45, 53]. Mittal et al. [45] utilized two network branches to learn from limited labeled and annotation-free samples; one branch was a GAN-based model. Souly et al. [53] proposed a GAN-based approach to train the network with additional weakly labeled data. They employed GAN to generate additional images useful for the classification task.

2.3 Contrastive Learning

Contrastive learning can be used in both supervised and unsupervised settings, and some studies have explored the application of contrastive learning on semantic segmentation [3, 11, 22, 34, 57, 60, 72, 73, 75]. The goal is to learn a representation space where positive pairs are close to each other whereas negative ones are far away. They utilize augmented versions of the same instance as positive pairs and others randomly sampled as negative ones. Some studies have investigated strategies to select negative pairs [11, 12, 22, 61]. Wu et al. [61] employed a memory bank to store representation and increase the number of negative samples. In MoCo [12, 22], a memory buffer and momentum encoder are proposed to build large and consistent dictionaries for unsupervised learning.

Recently, pixel-wise contrastive loss has been explored for semi-supervised semantic segmentation [3, 9, 34, 58, 73, 75]. Alonso et al. [3] and Zhou et al. [75] yielded intra-class similarity and inter-class discrimination when implementing contrastive loss. Wang et al. [58] proposed a fully supervised contrastive learning approach for semantic segmentation, emphasizing the similarity of embeddings within the same class and their dissimilarity across different classes. Chaitanya et al. [9] aimed to maximize intra-class similarity and inter-class separability for the segmentation task, which is effective for generic multi-class semantic segmentation. Our work specifically focuses on road extraction from images, which involves a binary classification task and differs from generic semantic segmentation. Our pixel-wise contrastive loss is specifically designed for the road extraction task, utilizing positive pairs of regions to capture contextual information and introducing a negative sampling strategy to identify valid negative region pairs.

The contrastive learning idea has also been exploited for remote sensing imagery classification in recent years [29, 32]. Jean et al. [29] introduced Tile2Vec, an unsupervised learning algorithm, for land cover classification and poverty prediction tasks. Their algorithm leverages the observation that spatially proximate remote sensing image tiles tend to exhibit similar representations, contrasting with those farther apart. By utilizing geospatial information as prior knowledge, Tile2Vec learns vector representations of remote sensing images. On the other hand, Kang et al. [32] proposed an unsupervised deep model for remote sensing imagery classification. They highlighted the semantic similarities among nearby geospatial locations and the diverse representations of contrasting land cover types. Although the above-mentioned contrastive learning methods have achieved promising results, they seldom considered pixel-wise contrastive loss with contextual information for road extraction tasks.

2.4 Self-training

Self-training is first applied for classification [8, 50, 67, 76] and has been commonly employed for semantic segmentation with deep learning [13, 19, 27, 45]. These techniques produce

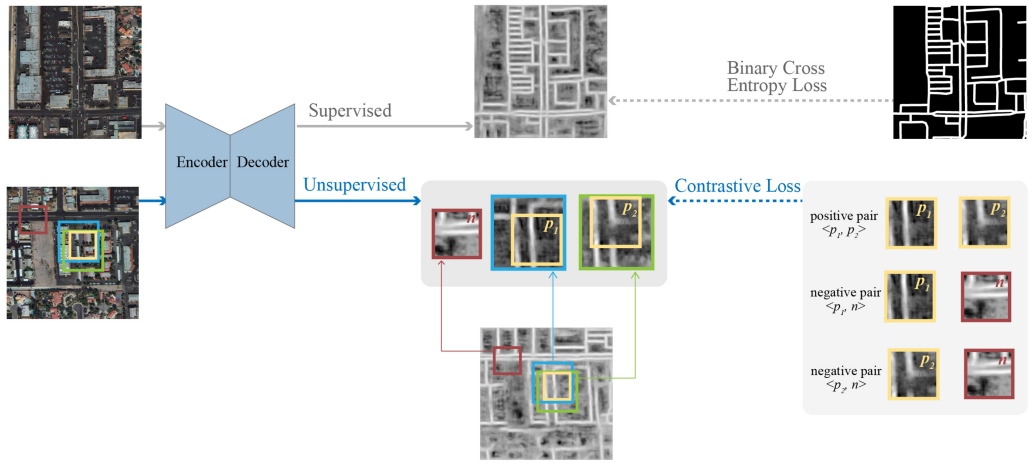

Fig. 3. Network Architecture. There are two network branches: one for labeled data and the other for unlabeled data. The supervised branch is trained with a binary cross entropy loss. The unsupervised branch is trained with the proposed pixel-wise contrastive loss defined over both positive pairs of regions $< p_1, p_2 >$ and negative pairs of regions $< p_1, n >$ or $< p_2, n >$. The blue and green boxes are two overlapping regions; their common regions (in yellow) form a positive pair. The red region and one of the yellow regions form a negative pair.

pseudo-segmented images by training a model with annotated data and utilizing the pseudo labels to retrain the models iteratively. Different strategies are designed to decide pseudo labels. For example, Cascante-Bonilla et al. [8] introduced curriculum learning to self-training, selecting pseudo labels with increasing thresholds. Hung et al. [26] proposed a GAN-based approach by designing a discriminator to distinguish the predicted confidence maps from the ground-truth segmentation distribution and select high-confident predicted segmentation as pseudo labels. In [13], cross pseudo supervision (CPS) was proposed, in which the pseudo segmentation labels produced by one network are exploited to supervise the other with perturbations and vice versa. Unlike the methods mentioned above, we propose a new iterative labeling strategy for self-training, which utilizes the proposed contrastive loss to select high-quality pseudo labels and increase the training dataset gradually. Our method is effective in road extraction with fewer labeled data.

3 METHOD

3.1 Method Overview

Figure 3 summarizes the sketch of the proposed semi-supervised method for extracting road regions in aerial images. The inputs to our method include a set of labeled images $\mathcal{D}_l = \{(x_l, y_l) \mid x_l \in X_l, y_l \in Y_l\}$ and a set of unlabeled images $\mathcal{D}_u = \{x_u \mid x_u \in X_u\}$, where (x_l, y_l) is a pair of image-label and x_u a raw image. The network includes two branches of sub-networks, one for supervised learning and the other for unsupervised learning. To train the supervised branch, we apply an encoder-decoder network over each training image x_l to extract its feature maps, and train the network using a binary cross-entropy loss $\mathcal{L}_S$.

The unsupervised branch shares the same network backbone as the supervised branch. It takes as input a raw image and is trained with the proposed contrastive loss function. The loss function is defined over both positive pairs of images and negative pairs. As Figure 3 illustrates, we crop two overlapping regions from the same image to form a positive pair. Similarly, two non-overlapping

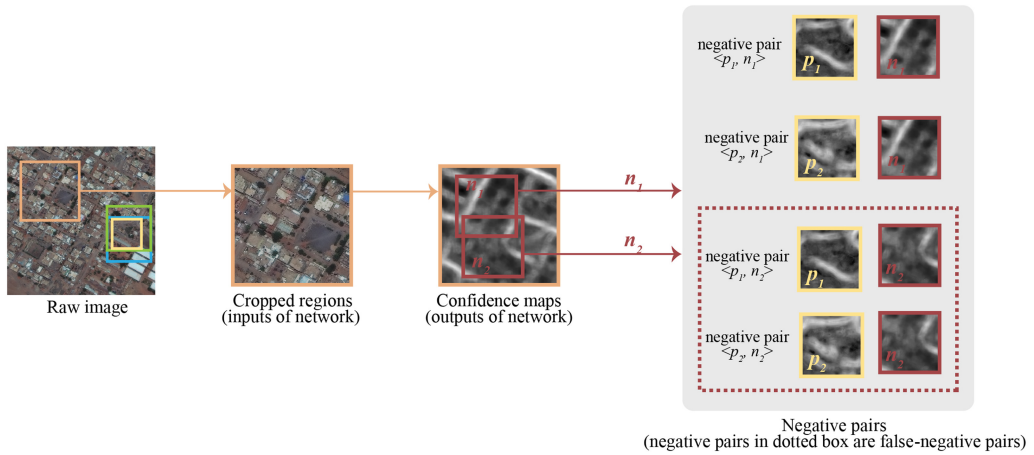

Fig. 4. Generation of negative pairs. For each positive pair of regions (green and blue boxes), we find and crop a non-overlapping region (orange box) from which a set of sub-regions ($n_1, n_2, ...$) are sampled and paired with the positive regions to form negative pairs. The proposed negative sampling algorithm is used to identify false-negative pairs (dotted box).

regions in the same image will be randomly selected and cropped to form a negative pair. Each cropped region is associated with a region in the network confidence map. The confidence maps over positive pairs of regions are expected to be similar, whereas the maps over negative pairs of regions are expected to be dissimilar. We employ this intuitive expectation to regularize the training of the network. To measure the similarities between the two confidence maps, we employ a histogram-based descriptor [16] to account for pixel-wise misalignment. A negative sampling strategy is incorporated to filter out false-negative samples in contrastive learning. We also develop an iterative process that employs contrastive loss to identify raw images with high-quality pseudo labels and retrains the network using both human labels and pseudo labels.

3.2 Pixel-wise Contrastive Loss

We introduce a pixel-wise contrastive loss function to supervise the training of the unsupervised network branch. This loss is defined over both positive and negative pairs of images. In this context, each pair of positive images denotes two overlapping regions; a pair of negative samples includes two image regions that do not overlap with each other. To collect a positive pair, we randomly select an image region and crop two image patches that enclose the overlapping region. Next, we select another region that is spatially far away from the above-shared region and pair it with the previously selected region to form a negative pair. The collected negative pairs might include false negatives, i.e., two disjoint regions yet sharing similar roadway patterns. We will introduce a sampling algorithm to remove these false-negative pairs. Figure 4 illustrates the process of generating negative pairs.

Let $< p_1, p_2 >$ denote a positive pair of image regions, $< p_1, n >$ or $< p_2, n >$ a negative pair, $\mathbf{sim}(p_1, p_2)$ returns the similarity between the network outputs of regions p_1 and p_2. One possible similarity measure is directly comparing the two network outputs pixel by pixel. However, it is not applicable here because there are misalignments caused by step-by-step downsampling network operations and contextual information fusion [25]. Instead, we extract the histogram of oriented gradients (HOG) from the network output of an image region and calculate the cosine similarity between HOG descriptors. The histogram-based features can ensure that the

ALGORITHM 1: Negative Sampling Strategy

Input: A pair of positive regions, p_i where $i = 1, 2$
Input: An image region having no overlap with the positive pair, q
Output: Valid negative samples, N_i where $i = 1, 2$

```
for i in [1, 2] do
    N = []
    j = 1
    while j ≤ 10 do
        sample a sub-region n_j from the region q to obtain a negative pair < p_i, n_j >
        N.append(n_j)
        j += 1
    end while
    N_i ← select 5 negative pairs with the lowest similarity score from N
end for
return N_i
```

measurement is invariant against geometric transformations, effective for the misalignment issue. The contrastive loss is defined as the following:

$$l(p_1, p_2, n) = -\log \frac{\exp(\mathbf{sim}(p_1, p_2)/\tau)}{\exp(\mathbf{sim}(p_1, p_2)/\tau) + \sum_n \exp(\mathbf{sim}(p_1, n)/\tau)} \tag{1}$$

$$\mathcal{L}_C = \sum_{<p_1, p_2, n>} (l(p_1, p_2, n) + l(p_2, p_1), n) \tag{2}$$

where τ is a temperature hyper-parameter [61]. Backpropagation of $l(p_1, p_2, n)$ is stopped for p_2. Note that each region in the positive pair might be paired with multiple non-overlapping regions to form different negative pairs. Then, the total semi-supervised loss $\mathcal{L}$ can be written as

$$\mathcal{L} = \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{L}_C, \tag{3}$$

where λ_1 and λ_2 are balancing weights for $\mathcal{L}_S$ and $\mathcal{L}_C$, respectively.

Negative Sampling Strategy. Filtering false-negative samples is a crucial component of the proposed contrastive learning because there is a high chance for two disjoint regions to have similar appearances. On the one hand, a large portion of a roadway aerial image might be of backgrounds; on the other hand, many road instances share similar shapes. Including these false-negative samples in the contrastive loss function would sacrifice the effectiveness of the training. Algorithm 1 and Figure 4 describe the proposed negative sampling process. We first apply the previously trained network to a cropped region that has no overlap with positive pairs to obtain its confidence map and HOG descriptor. Then, we sample sub-regions from the cropped region and calculate each negative pair's cosine similarity **sim**(). A negative pair with a high similarity score (or a large contrastive loss) would likely be a false negative. We sample 10 negative pairs and select 5 with the lowest similarity score.

3.3 Iterative Labeling

We develop an iterative labeling scheme to fully explore the set of unlabeled images and the proposed pixel-wise contrastive learning framework. Our iterative algorithm starts with initially training the network using both labeled data $\mathcal{D}_l$ and unlabeled data $\mathcal{D}_u$. Next, we apply the network

ALGORITHM 2: Iterative labeling

Input: Labeled data $\mathcal{D}_l$
Input: Unlabeled data $\mathcal{D}_u$
Input: Number of self-training iteration t
Input: Percentile threshold thr_1 (e.g., 80%)
Input: Contrastive loss threshold thr_2 (e.g., 4.7)
Output: Labels for self-training $[\mathcal{D}_1, \ldots, \mathcal{D}_{t-1}]$

$f(\cdot) \leftarrow$ train with $\mathcal{D}_l$
$i = 1$
while $i \leq t$ **do**
  $\mathcal{D}_i = \mathcal{D}_l$
  Rank all the raw images $\mathcal{D}_u$ and their confidence maps Sigmoid$(f(\mathcal{D}_u))$ based on average confidence scores
  $\mathcal{D}_m \leftarrow$ select the top ranked thr_1 raw images from $\mathcal{D}_u$
  Rank raw images $\mathcal{D}_m$ based on the contrastive loss
  $\mathcal{D}_s \leftarrow$ select raw images with contrastive loss smaller than thr_2 from $\mathcal{D}_m$ as pseudo labeled samples
  $\mathcal{D}_i = \mathcal{D}_i \cup \mathcal{D}_s$
  if $i > 1$ **then**
    $\mathcal{D}_i = \mathcal{D}_i \cup (\mathcal{D}_{i-1} \setminus \mathcal{D}_i)$
  end if
  $f(\cdot) \leftarrow$ train from scratch with $\mathcal{D}_i$
  $i = i + 1$
end while
return $[\mathcal{D}_1, \ldots, \mathcal{D}_{t-1}]$

over each raw image in $\mathcal{D}_u$ to obtain its confidence map. The estimated maps are used as the pseudo labels and will be used to enlarge the labeled data set $\mathcal{D}_l$ in the next iteration. We repeat the above process multiple times till the network converges. Similar iterative algorithms have been used in the relevant literature for various image tasks [8, 28, 69]. While being effective, a critical challenge to these iterative learning schemes is how to filter out low-quality pseudo labels to avoid model collapse. In this work, we employ the loss function to score pseudo labels and grow the labeled dataset. We crop overlapping regions from the pseudo labels to form the positive pair and use the negative sampling strategy to generate the negative pair. We compute the contrastive loss of pseudo labels with generated positive and negative pairs.

Algorithm 2 describes the sketch of the proposed iterative labeling process. At the beginning of the algorithm ($t = 1$), we select the top thr pseudo labels from Sigmoid$(f(\mathcal{D}_u))$ with the highest average confidence score and remove the images whose contrastive loss is above an empirical threshold. The contrastive loss describes how consistent the current network works for a raw image and can be used as a successful measure to select pseudo labels. We set the threshold to be 4.7 in this work. Figure 5 demonstrates some examples of raw images and pseudo labels with low (high-quality pseudo labels) and high contrastive loss (low-quality pseudo labels). We can combine these raw images with pseudo labels and the labeled set $\mathcal{D}_l$ to train the supervised branch. For the subsequent self-training ($t \geq 2$), we combine $\mathcal{D}_t$ and $\mathcal{D}_{t-1}$ together for the supervised branch. Compared with the original self-training with all Sigmoid$(f(\mathcal{D}_u))$ for training, our iterative labeling selection strategy can increase the quantity and quality of pseudo labels for the supervised branch and boost the model performance step by step.

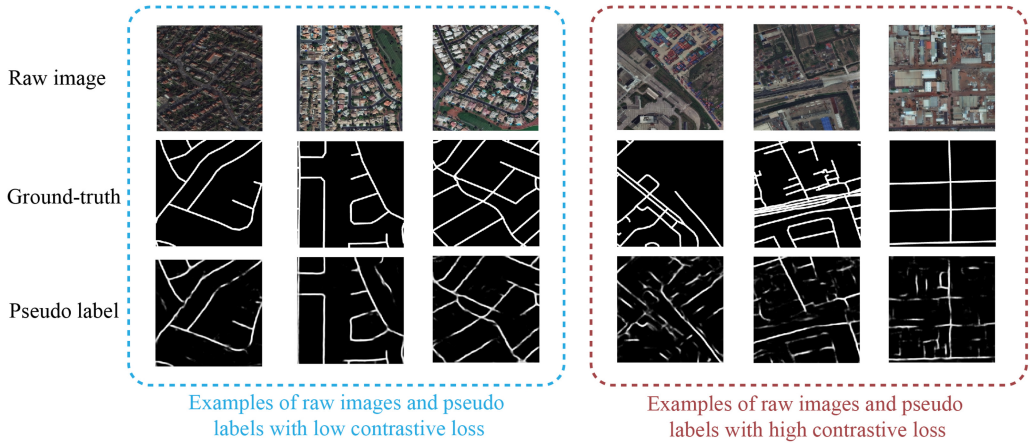

Fig. 5. Examples of raw images and pseudo labels with low contrastive loss (high-quality pseudo labels) and high contrastive loss (low-quality pseudo labels).

3.4 Training Data and Test-Time Augmentation

Training data augmentation (TA) is a common practice to produce training samples for supervised learning when data are insufficient, but an effective way is sought to produce different perturbations of the same unlabeled image [18, 36, 68, 74]. When training our semi-supervised model, we employ flipping, rotating, and color jitter augmentation for both supervised and unsupervised branches.

Test-time augmentation (TTA) contains augmentation, prediction, transformation back, and merging. For a test image x_t, we rotate it by 90°, 180°, and 270°. Then, we predict results based on both original and augmented images and transform the predicted results back. Towards the end, we average the four outputs to get the final prediction.

4 EXPERIMENTS AND RESULTS

4.1 Dataset

We apply the proposed semi-supervised approach to the public SpaceNet3 [56] and DeepGlobe Road [17] datasets and compare it to other alternative semi-supervised methods in the recent literature.

SpaceNet3. This dataset contains 2780 images. The size of each image is 1300 × 1300, with a ground resolution of 30 cm/pixel. We remove the images without roads and obtain a subset of 2549 images. The annotations are in the format of line strings for roads. Following [6], we first employ the Euclidean distance transform along the line strings to get the Gaussian maps and then threshold the map using a constant of 0.76 to obtain the binary mask, which corresponds to 6–7 meters wide road. We split the dataset into a subset of 2018 images for training, 100 for validation, and 431 for testing, following the approach described in [6].

DeepGlobe Road. This dataset includes a total of 6226 satellite images, which are captured over Thailand, Indonesia, and India, spanning 1632 km^2 in ground area. All pixels belong to the road areas. Each image has a size of 1024×1024 pixels and a ground resolution of 50 cm/pixel. Following [6, 51], we divide the dataset into 4496 for training, 200 for validation, and 1530 for testing.

4.2 Evaluation Metrics

We use the pixel-based *F1*-score and road intersection over union (*IoU*) to evaluate the model performance, following the previous works [1, 56]. The model outputs are confidence maps for

road extraction, and each threshold gives different results of *F1* and road *IoU*. We report the highest *F1* and road *IoU* as our evaluation metric. Similar evaluation metrics have been widely used in edge/boundary-related tasks in the computer vision community, such as ODS and OID at best scales in BSD500 [4, 63]. The motivation is to get the best performance for each model to compare and avoid evaluation bias, which ends up with different thresholds.

4.3 Implementations

We use U-Net [49] as the backbone network. For the supervised branch, we crop multiple image regions of 256×256 from each labeled image and use these cropped images to enlarge the training set. For the unlabeled branch, we first select a region of 256×256 and draw two outside regions of 384×384. These two larger regions overlap with each other, and their overlapping areas (256×256) form the positive pairs. To collect negative pairs, we crop a region of 512×512 having no overlap with positive pairs and select negative samples from this region that are perceptually dissimilar from positive samples. We extract a HOG descriptor from the network output over each region and measure their distance to prune similar regions. We also augment the cropped images by applying different transformations, including flipping, rotating, and color jitter augmentations.

We train all the models using a single GTX1080ti GPU with a batch size of 6 for labeled data and 2 for unlabeled data. Adam optimizer is implemented with a weight decay of 0.0001. The supervised and unsupervised loss weights are set as 1 and 0.1, respectively. There are a total of 100 epochs for SpaceNet3 and 60 for DeepGlobe Road. We start with a learning rate of 0.0005 with a step scheduler by a factor of 0.1 at epochs $\{60, 75\}$ for SpaceNet3 and $\{30, 40\}$ for DeepGlobe. The first four and two epochs are trained only with the supervised loss for stabilization for SpaceNet3 and DeepGlobe, respectively. The temperature hyper-parameter τ is 0.07. The patch size is 8×8, and the bin number is 12 to generate HOG descriptors. The percentile threshold thr_1 and contrastive loss threshold thr_2 are 80% and 4.7, respectively, to select pseudo labels for the iterative labeling process. The supervised baseline model is U-Net trained without TA or TTA. We compare our model on 1% and 5% of the labeled data of SpaceNet3 and DeepGlobe Road datasets.

4.4 Comparison with State-of-the-art Methods

We apply both the proposed method and other recently published semi-supervised segmentation methods, including C^3–SemiSeg [75], PC^{2}Seg [73], Cross Pseudo Supervise (CPS) [13], and Cross-Consistency Training (CCT) [47] over the same datasets and compare their performance. We compared models using the same partition protocols for fairness.

Results on SpaceNet3. Table 1 reports the comparison results of the proposed method and the other four methods on the SpaceNet3 dataset while using 1% or 5% of the labeled samples. We include the results of the baseline method for comparison. Besides the proposed iterative labeling method, we also implement a variant of our method that does not use the iterative labeling. The comparison in Table 1 suggests that the proposed semi-supervised method with iterative labeling (IL) achieves the highest performance in both settings and clearly outperforms the other four state-of-the-art semi-supervised segmentation methods [13, 47, 73, 75]. Moreover, the proposed method can still outperform other methods without iterative labeling. Note that there are only 20 training images using 1% labeled data. The proposed method can still work well and even outperforms the baseline method trained on (5%) labeled data.

Figure 6 further plots the F1 and IoU of both the supervised baseline and the proposed semi-supervised methods using 1%, 5%, and 100% labeled data. When we apply our method to the fully supervised setting in which all data are assigned for both supervised and unsupervised training, our method (*F1*: 0.7223; *IoU*: 0.5712) can still beat the baseline (*F1*: 0.7066; *IoU*: 0.5578) with an improvement of 1.57% and 1.34% for *F1* and road *IoU*, respectively.

Table 1. Performance on the SpaceNet3 Dataset Under Different Proportions of Labeled Samples

Method	1% (20)		5% (101)	
	F1	*IoU*	*F1*	*IoU*
C^3–SemiSeg [75]	0.5448	0.3749	0.5987	0.4279
PC^2Seg [73]	0.4788	0.3109	0.5601	0.3924
CPS [13]	0.3475	0.2002	0.5788	0.4037
CCT [47]	0.5785	0.4142	0.6358	0.4722
Sup. Baseline [49]	0.4886	0.3363	0.6052	0.4431
Ours w/o IL	0.6097	0.4480	0.6671	0.5038
Ours with IL	**0.6554**	**0.4900**	**0.6775**	**0.5044**

Ours w/o IL: our proposed semi-supervised learning of one single iteration without iterative labeling. Ours with IL: our proposed semi-supervised learning of multiple iterations of self-training with iterative labeling.

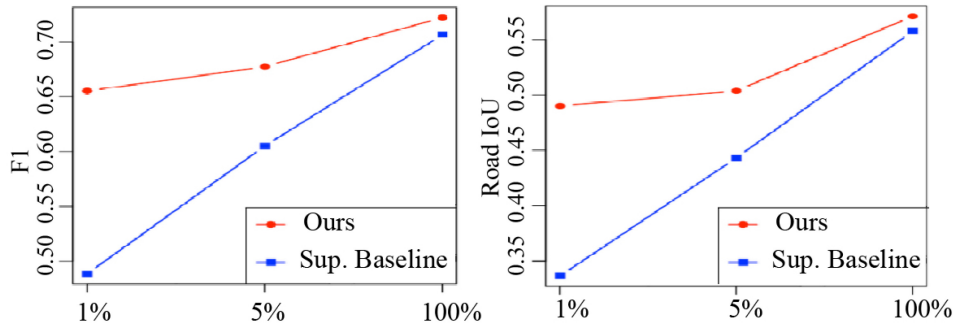

Fig. 6. Comparison between the supervised baseline model [49] and our proposed approach under varying proportions (1%, 5%, and 100%) of the labeled data on the SpaceNet3 dataset.

Figure 7 visualizes the results of our method and five other methods, including the supervised baseline and four state-of-the-art methods. For the images in the first seven rows, our method can produce much higher-quality results than others. The last three rows also show three challenging images for which our method did not work well. These images include complicated highways and partially visible roadways; these patterns did not appear in the training set. Actually, none of the six learning-based methods can work properly over these three images. One potential solution is the so-called sample-specific data augmentation method [36]. In our current experiment, we applied the same set of augmentation operations, including cropping, flipping, rotating, and color jitter, over every image. A more reasonable augmentation solution is to assess how difficult an image is regarding the current model, and to produce more augmented samples for these difficult samples than easy samples. We will explore this direction in future work.

Results on DeepGlobe Road. Table 2 reports the results of different methods over the DeepGlobe Road dataset. Some examples of results are shown in Figure 8. The recently proposed method C^3–SemiSeg [75] achieves F1 0.6383 while using 1% labeled data, which clearly outperforms other methods, including the baseline network. Our method with iterative labeling can further improve its performance with a significant margin of 3.46%. Similar improvements are obtained by the proposed method over other methods while using 5% labeled data. When we apply our method to the fully supervised setting in which all data are assigned for both supervised and unsupervised training, our method (*F1*: 0.7100; *IoU*: 0.5712) can still beat the baseline (*F1*: 0.6992; *IoU*: 0.5577).

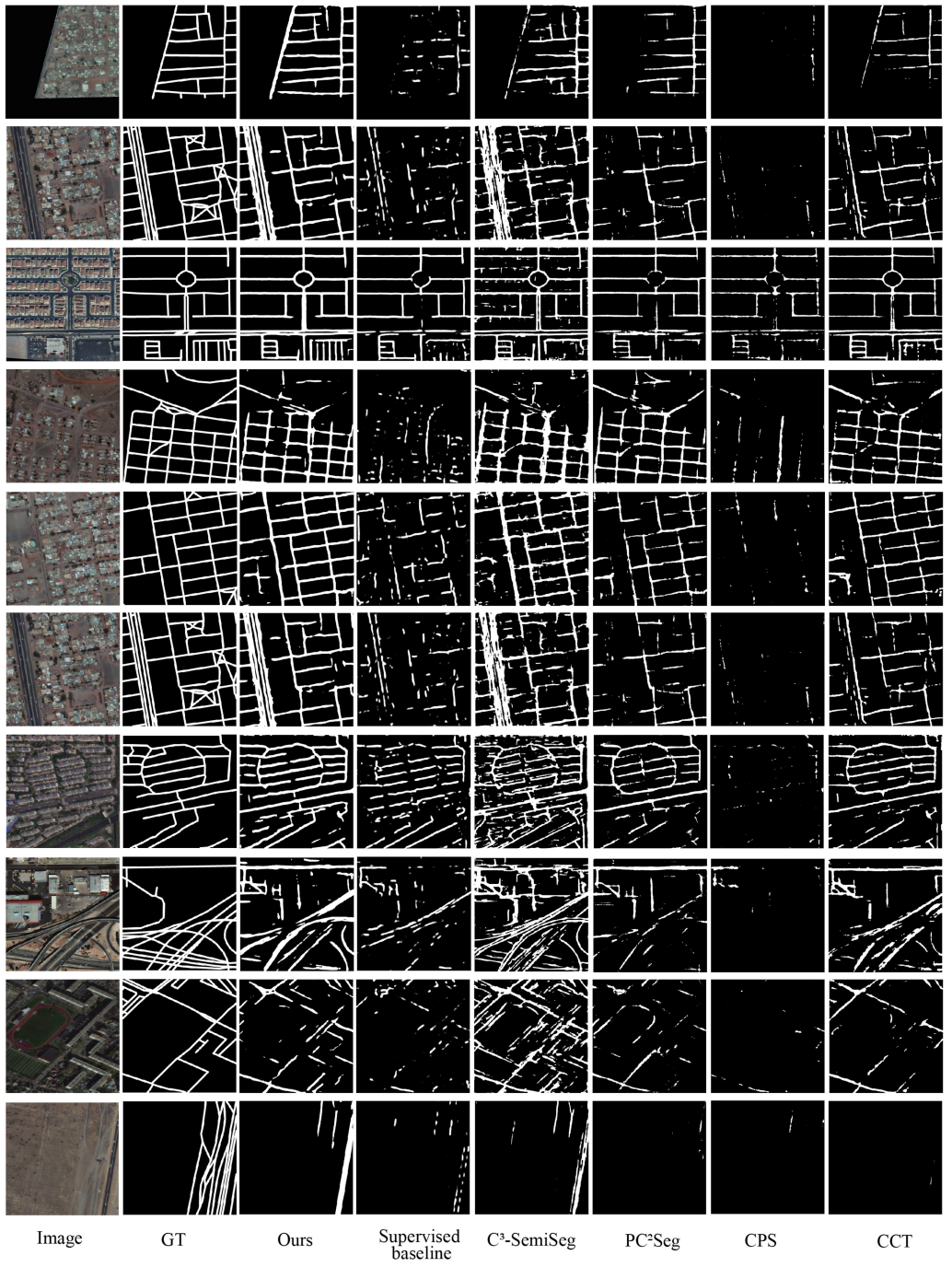

Fig. 7. Comparison results of six different methods on the SpaceNet3 dataset with 1% labeled data. The last three rows demonstrate some failure cases of our proposed method.

Table 2. Performance on the DeepGlobe Road Dataset Using 1% or 5% Labeled Samples

Method	1% (45)		5% (225)	
	F1	*IoU*	*F1*	*IoU*
C^3−SemiSeg [75]	0.6383	0.4530	0.6949	0.5232
PC^2Seg [73]	0.4693	0.2960	0.5987	0.4189
CPS [13]	0.2751	0.1549	0.5884	0.4079
CCT [47]	0.5475	0.3664	0.6305	0.4552
Sup. Baseline [49]	0.5763	0.3952	0.6800	0.5083
Ours w/o IL	0.6546	0.4745	0.7043	0.5352
Ours with IL	**0.6729**	**0.4956**	**0.7249**	**0.5605**

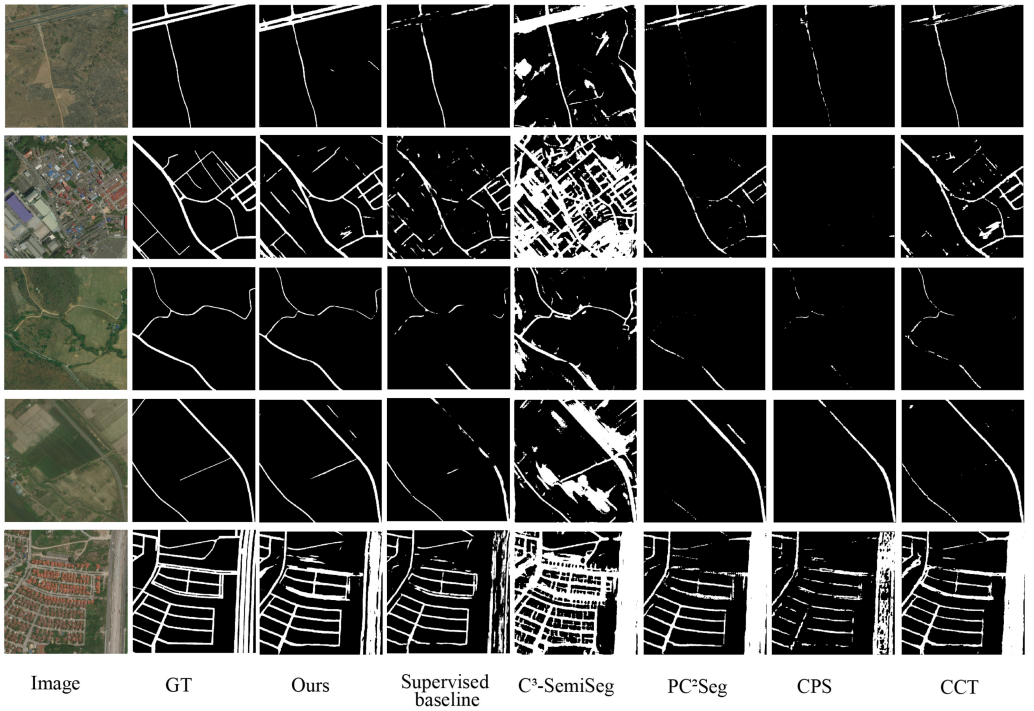

Fig. 8. Comparison results of six different methods on the DeepGlobe Road dataset with 1% labeled data.

4.5 Ablation Study

We perform ablation studies to validate the effectiveness of the proposed components on the SpaceNet3 dataset with 1% (i.e., 20) labeled and 99% (i.e., 1998) unlabeled training data. Table 3 reports the $F1$−score and road *IoU* of these methods. The method of ID I represents the supervised baseline model without training data augmentation (TA), test-time augmentation (TTA), pixel-wise contrastive loss (PCL), histogram of oriented gradients (HOG), negative sampling (NS), or self-training (ST).

Effectiveness of Histogram of Oriented Gradients (HOG). Misalignment is a challenge when working with pixel-wise tasks and calculating between-sample similarities. In this work, a roadway appears anywhere or at any scale in an image. We employ a histogram-based feature

Table 3. Ablation Study on the SpaceNet3 Dataset with 1% Labeled Data

	Model ID	TA	TTA	Context	PCL	HOG	NS	*F1*	*IoU*
Supervised	I							0.4886	0.3363
	II	✓						0.5507	0.3856
	III		✓					0.5223	0.3677
	IV	✓	✓					0.5791	0.4134
Semi-supervised	V			✓	✓	✓	✓	0.5604	0.4009
	VI		✓	✓	✓	✓	✓	0.5908	0.4206
	VII	✓		✓	✓	✓	✓	0.5914	0.4292
	VIII	✓	✓	✓	✓			0.5958	0.4342
	IX	✓	✓	✓	✓	✓		0.6028	0.4405
	X	✓	✓		✓	✓	✓	0.5896	0.4194
	XI	✓	✓	✓	✓	✓	✓	**0.6097**	**0.4480**

TA: training data augmentation; TTA: test-time augmentation; Context: contextual information; PCL: pixel-wise contrastive loss; HOG: histogram of oriented gradients; NS: negative sampling.

(HOG) to ensure that the contrastive loss is invariant against geometric transformations. In Table 3, the comparison between Model VIII and IX suggests that the semi-supervised model with HOG descriptors is superior to that without HOG, improving *F1* from 0.5958 to 0.6028 and road *IoU* from 0.4342 to 0.4405.

Effectiveness of Negative Sampling. The proposed negative sampling method aims to prune or filter false-negative pairs of samples during training. We evaluate and report the results of the semi-supervised model with and without negative sampling (NS). Models IX and XI results in Table 3 show that integrating NS can lead to improved *F1* and road *IoU*.

Effectiveness of Contextual Information. To verify the effectiveness of the contextual information on the predictions of overlap regions, we compare the models with and without contextual information. For the model without contextual information, we crop the same region with different data augmentations to generate positive pairs for the contrastive loss. The two cropped patches of the unlabeled image are totally overlapped and have the same contextual information. In Table 3, the comparison between Models X and XI shows that considering contextual information in the model can significantly improve F1 by 2.01% and IoU by 2.86%.

Effectiveness of Pixel-wise Contrastive Loss. Our proposed pixel-wise contrastive loss (PCL), considering contextual information, is performed on the pixel level. It encourages the network to have similar outputs for positive pairs and different results for negative ones. In Table 3, the comparisons between Model VIII and Model IV show that utilizing PCL over unsupervised images can lead to a gain of 1.67% in *F1* and 2.08% in road *IoU* while using augmentation techniques for both testing and training. We observe an even more significant improvement from the comparisons of Model V and Model I. Using PCL and negative sampling will lead to a gain of 7.18% in *F1* and 6.46% in road *IoU*. These experiments with comparisons indicate that our PCL is effective for road extraction as it helps to learn more semantic representations for the model and alleviate the over-fitting issue.

Effectiveness of Iterative Labeling. Table 4 reports how the proposed method works over iterations. The method called *plain ST w/o IL* is a naïve implementation of the proposed method in which the pseudo labels at the previous iterations are directly used as the labeled samples at the current iteration. When applying this naïve implementation with our semi-supervised model, we observe that the *F1* and road *IoU* decrease in the second iteration. In contrast, the method called *ST with IL*, which uses the proposed iterative labeling strategy, can progressively improve the *F1* and road *IoU* of our semi-supervised model over the first four iterations and converge at the

Table 4. Effectiveness of Iterative Labeling on the SpaceNet3 Dataset with 1% Labeled Data

Method	t	*F1*	*IoU*
Supervised baseline with plain ST w/o IL	0	0.4886	0.3363
	1	0.5375	0.3791
	2	0.5491	0.3906
	3	**0.5597**	**0.3991**
Our proposed semi-supervised with TA, TTA, PCL, HOG, NS, and plain ST w/o IL	0	0.6097	0.4480
	1	**0.6399**	**0.4740**
	2	0.6368	0.4637
Our proposed semi-supervised with TA, TTA, PCL, HOG, NS and ST with IL	0	0.6097	0.4480
	1	0.6425	0.4779
	2	0.6482	0.4837
	3	0.6534	0.4870
	4	**0.6554**	**0.4900**
	5	0.6551	0.4872

ST: self-training; IL: iterative labeling; t: iteration of self-training model. See text for more details.

fifth iteration. The best performance of our semi-supervised iterative approach on the 1% labeled SpaceNet3 dataset achieves 0.6551 *F1* and 0.4900 road *IoU*, which are much higher than those of the supervised baseline model employing plain ST w/o IL. This set of comparisons shows that the proposed iterative labeling process can effectively prune the noisy data.

In this study, we employ contrastive loss during the iterative labeling process to filter out low-quality pseudo labels. Our offline experiments show that directly ranking all pseudo-labels based on their contrastive loss and classification loss can slightly enhance system robustness. In previous literature, various measures have been introduced to identify and select high-quality pseudo-labels when learning from raw data. These measures include confidence scores, diversity of selected samples, and other learning-based metrics. However, finding an optimal combination of these measures is a non-trivial problem. In this study, we focus exclusively on the proposed pixel-wise constructive loss and empirically demonstrate its effectiveness as a measure for selecting pseudo-labels. We remain open to exploring different measurement combinations to select high-quality pseudo labels in future research.

Effectiveness of Training Data and Test-Time Augmentation. As shown in Table 3, the supervised and semi-supervised models with TA (Model II and Model VII) have better *F1* and road *IoU* than corresponding models without TA (Model I and Model V). Model III and Model VI with TTA also perform better than Model I and Model V, respectively. When both TA and TTA are incorporated together in the supervised Model IV and semi-supervised Model XI, we obtain better results than using TA or TTA alone. These comparisons suggest that integrating TA, TTA, or both can significantly improve the model performance of road extraction.

Analysis on the 1% Labeled DeepGlobe Road Dataset. We further validate the effectiveness of the proposed components on the DeepGlobe Road dataset with 1% (i.e., 45) labeled and 99% (i.e., 4451) unlabeled data. Table 5 reports the results of multiple implementations of our methods. We can observe consistent improvements while using TA & TTA, PCL & HOG & NS, ST with IL, or the combinations. The model incorporating ST with IL achieves the best performance, whose *F1* is 9.66% higher and road *IoU* is 10.04 higher than the baseline method.

Analysis on the 5% Labeled SpaceNet3 Dataset. Table 6 reports the performance of our model on the SpaceNet3 dataset while using 5% (i.e., 101) labeled and 95% (i.e., 1917) unlabeled

Table 5. Effectiveness of Our Proposed Components on the DeepGlobe Road Dataset with 1% Labeled Data

TA& TTA	PCL& HOG&NS	ST with IL	*F1*	*IoU*
			0.6052	0.4431
✓			0.6374	0.4736
	✓		0.6304	0.4676
✓	✓		0.6671	0.5038
✓	✓	✓	**0.6775**	**0.5044**

TA: training data augmentation; TTA: test-time augmentation; PCL: pixel-wise contrastive loss; HOG: histogram of oriented gradients; NS: negative sampling; ST: self-training; IL: iterative labeling.

Table 6. Effectiveness of Our Proposed Components on the SpaceNet3 Dataset with 5% Labeled Data

TA& TTA	PCL& HOG&NS	ST with IL	*F1*	*IoU*
			0.5763	0.3952
✓			0.6496	0.4696
	✓		0.5944	0.4132
✓	✓		0.6546	0.4745
✓	✓	✓	**0.6729**	**0.4956**

data. TA & TTA, PCL & HOG & NS, and ST with IL effectively boost model performance, which is consistent with our previous experimental results obtained from 1% labeled SpaceNet3 data. The F1 score of our proposed semi-supervised iterative method achieves 0.6774, which outperforms the result of the supervised model trained with 10% labeled data (0.6295) and is only 2.9% lower than the supervised model trained over the full dataset (0.7066).

5 CONCLUSION

In this article, we propose an iterative semi-supervised approach for extracting roads in aerial images. We developed a pixel-wise contrastive loss to force similarity between positive pairs and dissimilarity for negative pairs. An iterative labeling scheme is employed to fully explore the knowledge of raw unlabeled images and generate pseudo image labels. Based on the proposed contrastive loss, a negative sampling strategy is developed to filter false-negative samples. Extensive experiments on public datasets demonstrate that our method achieves state-of-the-art in the task of image-based road extraction and outperforms the other semi-supervised segmentation models. Our method can also be generalized well on different datasets while trained on varying amounts of labeled data. The developed techniques have broad applications in multimedia computing, including video parsing, image understanding, and classification.

REFERENCES

[1] Abolfazl Abdollahi, Biswajeet Pradhan, Nagesh Shukla, Subrata Chakraborty, and Abdullah Alamri. 2020. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sensing* 12, 9 (2020), 1444.

[2] Roohallah Alizadehsani, Danial Sharifrazi, Navid Hoseini Izadi, Javad Hassannataj Joloudari, Afshin Shoeibi, Juan M. Gorriz, Sadiq Hussain, Juan E. Arco, Zahra Alizadeh Sani, Fahime Khozeimeh, et al. 2021. Uncertainty-aware

semi-supervised method using large unlabeled and limited labeled COVID-19 data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 3s (2021), 1–24.

[3] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C. Murillo. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8219–8228.

[4] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2010), 898–916.

[5] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. 2018. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4720–4728.

[6] Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, C. V. Jawahar, and Manohar Paluri. 2019. Improved road connectivity by joint learning of orientation and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10385–10393.

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. 41–48.

[8] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2020. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001* (2020).

[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis* 87 (2023), 102792.

[10] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. 2020. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*. Springer, 695–714.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. PMLR, 1597–1607.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[13] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2613–2622.

[14] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. 2017. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing* 55, 6 (2017), 3322–3337.

[15] Jifeng Dai, Kaiming He, and Jian Sun. 2015. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1635–1643.

[16] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

[17] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 2018. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 172–181.

[18] Terrance DeVries and Graham W. Taylor. 2017. Improved regularization of convolutional neural networks with Cutout. *arXiv preprint arXiv:1708.04552* (2017).

[19] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. 2020. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514* 1, 2 (2020), 5.

[20] Geoffrey French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. 2019. Consistency regularization and CutMix for semi-supervised semantic segmentation. *arXiv preprint arXiv:1906.01916* 2, 4 (2019), 5.

[21] Geoff French, S. Laine, Timo Aila, Michal Mackiewicz, and G. Finlayson. 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv: Computer Vision and Pattern Recognition* (2020).

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[23] Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M. Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. 2020. Sat2Graph: Road graph extraction through graph-tensor encoding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 51–67.

[24] Namdar Homayounfar, Wei-Chiu Ma, Justin Liang, Xinyu Wu, Jack Fan, and Raquel Urtasun. 2019. Dagmapper: Learning to map by discovering lane topology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2911–2920.

[25] Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S. Huang, and Humphrey Shi. 2021. AlignSeg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1 (2021), 550–557.

[26] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. 2018. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934* (2018).

[27] Mostafa S. Ibrahim, Arash Vahdat, Mani Ranjbar, and William G. Macready. 2020. Semi-supervised semantic image segmentation with self-correcting networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12715–12725.

[28] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5070–5079.

[29] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3967–3974.

[30] John R. Jensen and Dave C. Cowen. 1999. Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing* 65 (1999), 611–622.

[31] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. 2019. Consistency-based semi-supervised learning for object detection. *Advances in Neural Information Processing Systems* 32 (2019).

[32] Jian Kang, Ruben Fernandez-Beltran, Puhong Duan, Sicong Liu, and Antonio J. Plaza. 2020. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing* 59, 3 (2020), 2598–2610.

[33] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. 2020. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647* (2020).

[34] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1205–1214.

[35] Lerenhan Li, Yunlong Dong, Wenqi Ren, Jinshan Pan, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. 2019. Semi-supervised image dehazing. *IEEE Transactions on Image Processing* 29 (2019), 2766–2779.

[36] Pu Li, Xiaobai Liu, and Xiaohui Xie. 2021. Learning sample-specific policies for sequential image augmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4491–4500.

[37] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. 2019. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems* 32 (2019).

[38] Ruyi Liu, Qiguang Miao, Yi Zhang, Maoguo Gong, and Pengfei Xu. 2019. A semi-supervised high-level feature selection framework for road centerline extraction. *IEEE Geoscience and Remote Sensing Letters* 17, 5 (2019), 894–898.

[39] Xiaoming Liu, Shuo Wang, Ying Zhang, and Quan Yuan. 2022. Scribble-supervised meibomian glands segmentation in infrared images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 3 (2022), 1–23.

[40] Xiaoyan Lu, Yanfei Zhong, Zhuo Zheng, Yanfei Liu, Ji Zhao, Ailong Ma, and Jie Yang. 2019. Multi-scale and multi-task deep learning framework for automatic road extraction. *IEEE Transactions on Geoscience and Remote Sensing* 57, 11 (2019), 9362–9377.

[41] Zhen Lv, Yonghong Jia, Qian Zhang, and Yifu Chen. 2017. An adaptive multifeature sparsity-based model for semiautomatic road extraction from high-resolution satellite images in urban areas. *IEEE Geoscience and Remote Sensing Letters* 14, 8 (2017), 1238–1242.

[42] Mehdi Maboudi, Jalal Amini, Shirin Malihi, and Michael Hahn. 2018. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. *ISPRS Journal of Photogrammetry and Remote Sensing* 138 (2018), 151–163.

[43] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. 2017. DeepRoadMapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*. 3438–3446.

[44] Zelang Miao, Wenzhong Shi, Paolo Gamba, and Zhongbin Li. 2015. An object-based method for road network extraction in VHR satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, 10 (2015), 4853–4862.

[45] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. 2019. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[46] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. 2018. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3136–3145.

[47] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12674–12684.

[48] Hong Pan, Yonghong Jia, and Zhen Lv. 2018. An adaptive multifeature method for semiautomatic road extraction from high-resolution stereo mapping satellite images. *IEEE Geoscience and Remote Sensing Letters* 16, 2 (2018), 201–205.
[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.
[50] Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.
[51] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar. 2018. Self-supervised feature learning for semantic segmentation of overhead imagery. In *BMVC*, Vol. 1. 4.
[52] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757* (2020).
[53] Nasim Souly, Concetto Spampinato, and Mubarak Shah. 2017. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*. 5688–5696.
[54] Ziyi Sun, Yunfeng Zhang, Fangxun Bao, Ping Wang, Xunxiang Yao, and Caiming Zhang. 2022. SADnet: Semi-supervised single image dehazing method based on an attention mechanism. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–23.
[55] Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. 2020. Vecroad: Point-based iterative graph exploration for road graphs extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8910–8918.
[56] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. 2018. SpaceNet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232* (2018).
[57] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 943–952.
[58] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. 2021. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7303–7313.
[59] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. 2018. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7268–7277.
[60] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. 2021. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10551–10560.
[61] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3733–3742.
[62] Junsheng Xiao, Huahu Xu, Honghao Gao, Minjie Bian, and Yang Li. 2021. A weakly supervised semantic segmentation network by aggregating seed cues: The multi-object proposal generation perspective. *ACM Transactions on Multimedia Computing Communications and Applications* 17, 1s (2021), 1–19.
[63] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 1395–1403.
[64] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3060–3069.
[65] Xinxing Xu, Wen Li, Dong Xu, and Ivor W. Tsang. 2015. Co-labeling for multi-view weakly labeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 6 (2015), 1113–1125.
[66] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).
[67] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3 (2017), 1–23.
[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6023–6032.
[69] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems* 34 (2021), 18408–18419.

[70] Yang Zhang, Zhangyue Xiong, Yu Zang, Cheng Wang, Jonathan Li, and Xiang Li. 2019. Topology-aware road network extraction via multi-supervised generative adversarial networks. *Remote Sensing* 11, 9 (2019), 1017.

[71] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. 2018. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters* 15, 5 (2018), 749–753.

[72] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. 2021. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10623–10633.

[73] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. 2021. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7273–7282.

[74] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13001–13008.

[75] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. 2021. C3-SemiSeg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7036–7045.

[76] Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3, 1 (2009), 1–130.

Received 28 September 2022; revised 20 May 2023; accepted 21 June 2023