DualStrip-Net: A Strip-based Unified Framework for Weakly- and Semi-Supervised Road Segmentation from Satellite Images

Jingtao Hu, Qiang Li, *Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Automated road segmentation from remote sensing imagery remains a fundamental challenge in Earth observation systems. The primary bottleneck lies in acquiring dense pixel-wise annotations, which is both labor-intensive and time-prohibitive. This paper presents DualStrip-Net, a novel deep learning framework for weakly-supervised and semi-supervised road segmentation that effectively handles both sparse annotations and limited labeled data. Unlike conventional CNN-based segmentation methods that lack explicit road topology modeling, DualStrip-Net exploits the inherent linear topology of road networks through a dual-stream architecture that combines patch-level annotation strategy and strip-based feature learning. The framework captures road characteristics through orthogonal strip processing in horizontal and vertical orientations. The proposed DualStrip Learning mechanism enables robust feature representation of road structures through complementary views. Extensive evaluations on the DeepGlobe, Massachusetts, and CHN6-CUG benchmark datasets demonstrate that DualStrip-Net achieves superior performance in both weakly-supervised and semi-supervised settings. Notably, with only 20% of labeled training data, our method outperforms the supervised-only baselines on both Massachusetts and CHN6-CUG datasets. Code is available at https://github.com/jasonnhu/DualStrip-Net/.**

***Index Terms*—Remote sensing imagery, road segmentation, weakly-supervised, semi-supervised, patch-level, dualstrip.**

I. Introduction

Earth observation through remote sensing has become increasingly essential for understanding and managing our rapidly evolving planet [1]–[3]. In particular, accurate road extraction from satellite imagery plays a critical role in urban planning, navigation systems, and disaster response [4], [5]. While the exponential growth in Earth observation data offers unprecedented opportunities for automated road mapping, the inherent complexity of road networks in remote sensing imagery presents numerous technical challenges: varying road widths, diverse appearances under different conditions, frequent occlusions by buildings and vegetation, and complex topological structures at intersections [6].

While traditional fully supervised methods demonstrate promising results, they require extensive pixel-level annotations that are prohibitively time-consuming and labor-intensive, especially for large-scale mapping tasks. To mitigate this annotation burden, researchers explore two complementary directions. The first direction leverages weakly-supervised learning with more cost-effective annotation strategies. As shown in Fig. 1(a), these approaches range from scribble-level supervision [7]–[10] where experts trace road centerlines for approximate spatial location and topological connectivity guidance, to point-level annotations [11], [12] that require only sparse clicks (approximately 60s per image [11]), and image-level labels [13] that merely indicate road presence.

The second direction employs semi-supervised learning to harness abundant unlabeled data, thereby complementing the weakly-supervised approaches. Recent advances in semantic segmentation have demonstrated promising results through various consistency regularization strategies. For example, cross pseudo supervision (CPS) [14] enforces consistency between two networks with different initializations, where the pseudo labels from one network supervise the other network. The advanced self-training framework (ST++) [15] introduces multi-stage training with curriculum learning. More recent approaches such as FixMatch [16] and UniMatch [17] advance

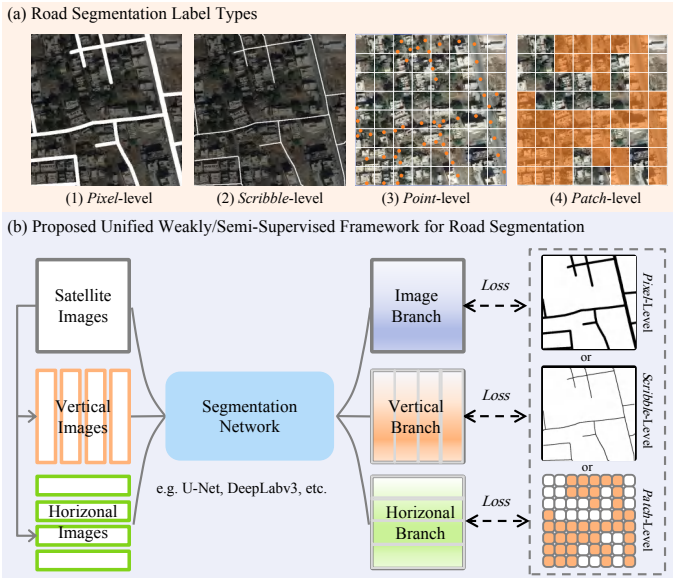

Fig. 1. Overview of our proposed DualStrip-Net framework. (a) Different types of supervision signals for road segmentation, including pixel-level, scribble-level, point-level, and our proposed patch-level annotations. (b) Proposed unified framework for both weakly-supervised and semi-supervised learning with a flexible architecture to incorporate various types of Labels.

This work was supported by the National Natural Science Foundation of China under Grant 62301385, 62471394, and U21B2041.

Jingtao Hu is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China. (e-mail: jthu@mail.nwpu.edu.cn). (*Corresponding author: Qi Wang.*)

Qiang Li and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: liqmges@gmail.com,crabwq@gmail.com).

this direction by introducing weak-to-strong consistency regularization. When adapted to road segmentation by methods like SemiRoadExNet [18] and MCMCNet [19], these principles exhibit limitations in preserving the inherent characteristics of road networks. This discrepancy between general semantic segmentation and road-specific requirements underpins our research.

To address these limitations, researchers have explored various approaches to enhance feature learning and structural consistency. In the weakly-supervised domain, Class Activation Mapping (CAM) [20] based methods have emerged as a potential solution, leveraging image-level supervision to generate coarse localization maps. Nevertheless, existing methods encounter several fundamental limitations: 1) CAM-based methods often struggle with fine-grained road details due to the coarse nature of activation maps; 2) point and scribble annotations, while efficient to collect, provide limited information about road width and boundaries, leading to inaccurate road surface delineation; and 3) current semi-supervised approaches lack effective mechanisms to maintain consistent road features across different views and scales. These challenges are particularly pronounced in remote sensing imagery, where road networks exhibit diverse widths and complex intersections.

Based on road network characteristics, we observe that roads exhibit strong structural continuity in both horizontal and vertical directions [21]. This property suggests a feature capture method that mirrors human perception of road structures [4]. Motivated by this observation, we propose a patch-level annotation strategy that requires only binary labels for road presence in regular image patches. This strategy reduces annotation burden while preserving structural information. Inspired by puzzle-like learning in weakly-supervised object localization [22], we develop DualStrip-Net (Fig. 1(b)) that processes images through complementary horizontal and vertical strips. Our framework explicitly models road continuity through complementary views and provides spatial constraints for supervision signal propagation. Furthermore, it enables natural consistency regularization between labeled and unlabeled data through strip-based feature matching.

Our main contributions can be summarized as follows:

- We propose an efficient patch-level annotation strategy that requires only binary labels for road presence, which provides effective supervision signals while reducing annotation effort.
- We present a unified framework, DualStrip-Net, that effectively handles both weakly-supervised and semi-supervised road segmentations through complementary strip-based learning.
- We conduct comprehensive experiments on three challenging datasets that demonstrate superior performance in both weakly- and semi-supervised settings.

The remainder of this paper is organized as follows. Section II reviews related work on road segmentation and learning with limited annotations. Section III presents our proposed DualStrip-Net framework. Section IV provides experimental results and analysis. Finally, Section V concludes the paper.

II. Related Work

Extracting accurate and up-to-date road networks from high-resolution satellite imagery remains a fundamental yet challenging problem in remote sensing and computer vision. Conventional fully supervised road segmentation methods show promising results in various scenarios [21], [23]–[26]. However, such methods are heavily dependent on pixel-level annotated training data, which is both expensive and time-consuming to produce. To alleviate the reliance on dense and accurate ground-truth labels, recent research efforts shift toward weakly supervised and semi-supervised frameworks to learn road extraction models under uncompleted and limited annotations.

A. Weakly-supervised Road Segmentation

Weakly supervised road segmentation typically relies on less demanding annotations, such as scribbles [8], [27], [28], sparse points [11], or existing cartographic data with incomplete or imprecise road information [29], [30]. For example, Wei et al. [8] propose a scribble-based weakly supervised deep learning method for road surface extraction, which utilizes centerlines as sparse supervision and a label propagation algorithm to generate proposal masks. Similarly, Lian and Huang [11] introduce a point-based weakly supervised approach that leverages only a few annotated pixels per image, further refined by machine learning classifiers and energy-based active contours. Additionally, Wu et al. [29] employ OpenStreetMap centerlines as weak annotations. They refine coarse labels with super-pixel segmentation and spectral-spatial priors to build weakly supervised convolutional models to reduce dependency on pixel-accurate labels. Recent advances also include framework improvements through contrastive and structure-aware learning. For instance, Yuan et al. [27] integrate adversarial learning into scribble-supervised methods to refine the pseudo-label quality and enhance the invariance of road segmentation predictions.

Weakly supervised methods face two main challenges: annotation incompleteness and label quality uncertainty. To address these, researchers explore auxiliary information such as boundary priors [31], data-driven invariance learning [28], and advanced refinement techniques to bridge the gap between weak supervision and fully annotated conditions. These methods often incorporate geospatial constraints, multi-scale features, and color/spatial clustering to guide the label propagation from sparse annotations to the entire scene.

B. Semi-supervised Road Segmentation

Semi-supervised methods combine a limited amount of pixel-level ground truth with a large corpus of unlabeled or weakly labeled images through consistency regularization and pseudo-labeling strategies. Chen et al. [30] introduce SW-GAN, an adversarial framework that utilizes a small set of fully annotated data and abundant weakly annotated data to train reliable road extraction models. Similarly, Zhang et al. [32] propose iterative self-training with pixel-wise contrastive losses to refine pseudo-labels and improve model performance.

Numerous semi-supervised learning (SSL) strategies exist in the context of road segmentation. Consistency regularization, which enforces prediction invariance across transformations, is one of the most widely adopted approaches across studies [33]–[35]. For example, Yang et al. [33] introduce a multi-scale consistency framework with a hybrid CNN-Transformer architecture (SSEANet) to achieve robust segmentation performance through local feature extraction and global attention mechanisms. Pseudo-labeling, another dominant SSL technique, generates pseudo-labels for unlabeled samples through contrastive learning modules [19], [36] or iterative refinements [36]. Gao et al. [19] enhance pseudo-labeling quality by integrating guided contrastive learning and multitask heads to improve road connectivity through skeleton predictions.

To address road-specific challenges, methods that focus on edge and connectivity show significant progress. Methods like MCMCNet [19] and SSEANet [33] emphasize edge awareness through multiscale features and auxiliary decoding modules to boost connectivity, especially for thin and occluded roads. Zhang et al. [32] propose a novel approach with pixel-wise contrastive loss to refine pseudo-labels and improve segmentation performance on narrow road structures. Additionally, other studies apply multi-scale consistency constraints to handle varying road widths and scales in optical remote sensing imagery [5], [19], [33].

Recent architectural innovations in SSL methods for road segmentation focus on hybrid designs. For example, models combining Convolutional Neural Networks (CNNs) and Transformers [33], [37] effectively balance local feature extraction and long-range attention to improve road topology extraction across scales and occlusions. Additionally, Meng et al. [38] demonstrate that large-scale pre-training with OSM data significantly improves model generalization across different regions. These semi-supervised algorithms often incorporate strong priors about road topology and geometric continuity to produce more robust and topologically coherent road networks than their fully supervised counterparts trained on limited data.

In contrast to existing approaches, we present a novel patch-level annotation strategy integrated with a dualstrip learning mechanism. The proposed framework addresses both weakly-supervised and semi-supervised scenarios and preserves road structural integrity through complementary views. This approach overcomes the limitations of existing methods through strip-based learning that leverages the inherent linear and continuous characteristics of roads and combines feature consistency constraints with structural preservation mechanisms in both weakly-supervised and semi-supervised settings.

III. Method

In this section, we present DualStrip-Net, a novel framework that unifies weakly-supervised and semi-supervised road segmentation through complementary strip-based learning. Given a remote sensing dataset $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$, where $\mathcal{D}_l$ contains images with different levels of supervision (pixel-wise, scribble-level, and patch-level binary labels) and $\mathcal{D}_u$ represents unlabeled images, our goal is to learn a robust road segmentation model that effectively leverages both types of supervision while preserving road structural properties.

A. Motivation

Road extraction from remote sensing imagery has made significant progress with WSL methods. Scribble-based approaches [8], which provide explicit road topology guidance through centerline annotations, have shown promising results in road segmentation. However, as shown in Fig. 4, these methods face inherent limitations in regions where road width varies significantly, particularly in transitions between highways and local roads. In contrast, patch-level binary labels, which only require annotators to mark the presence or absence of roads in local regions, offer a more efficient annotation strategy. Patch-level supervision provides less explicit topological information yet encourages the model to learn the complete set of road features within each region, rather than focusing primarily on centerline characteristics.

SSL methods, which leverage abundant unlabeled data to enhance model performance [17], [30], have emerged as a promising direction. However, directly applying these methods to road segmentation overlooks the unique topological properties of road networks, such as their continuous strip-like structures, strong directional patterns, and varying intersection complexities. These intrinsic characteristics require specialized architectural designs that can effectively capture and preserve road topology during feature learning. This observation motivates our DualStrip-Net framework that incorporates complementary horizontal and vertical views, which explicitly model road characteristics and effectively address both weakly- and semi-supervised road segmentation through strip-based feature learning.

B. DualStrip-Net

1) DualStrip Learning: The core idea of DualStrip learning is to exploit the inherent strip-like structure of roads through horizontal and vertical strip supervision. Given an input remote sensing image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, we decompose it into directional strips,

$$\mathcal{P}(x,p) = \begin{cases} \mathbf{x}_h, & \text{if } p = 0 \text{ (horizontal)}, \\ \mathbf{x}_v, & \text{if } p = 1 \text{ (vertical)}, \end{cases} \tag{1}$$

where $\mathcal{P}$ is the strip decomposition function. The strip generation process follows three key principles: 1) Structural Alignment: Strips are oriented along orthogonal horizontal and vertical directions not because roads follow these specific orientations, but because this complementary decomposition creates a complete representation that can effectively capture road structures at any orientation; 2) Coverage Balance: The strip size (determined by m and n) should balance between local detail preservation and global context modeling, as demonstrated in our ablation studies (Section IV-F); 3) Boundary Awareness: Strips should account for road continuity across boundaries, which is later addressed by our Dynamic Boundary-padded Cutting strategy. The horizontal and vertical strip sets are defined as:

$$\mathbf{x}_h = \{\mathbf{x}_{i,j} \in \mathbb{R}^{3\times h\times w} | i \in [0, n-1], j \in [0, m-1]\}, \tag{2}$$

$$\mathbf{x}_v = \{\mathbf{x}_{i,j} \in \mathbb{R}^{3\times h\times w} | i \in [0, m-1], j \in [0, n-1]\}, \tag{3}$$

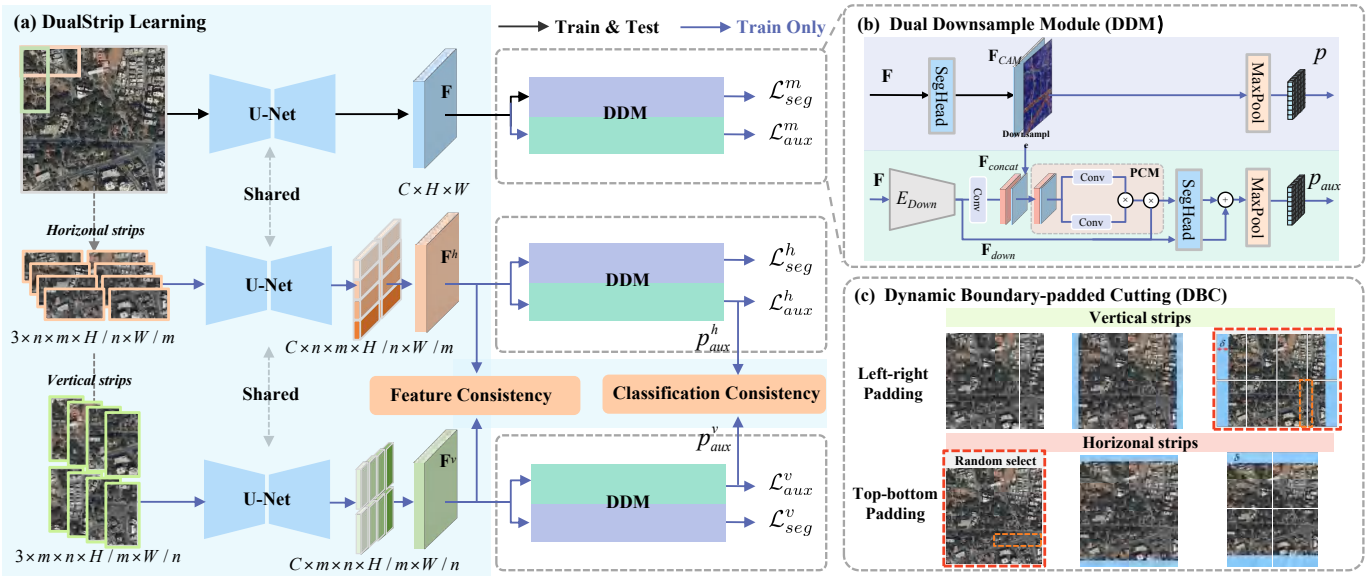

Fig. 2. Overview of our DualStrip-Net for weakly-supervised road segmentation. The framework consists of three key components: a) DualStrip Learning that enforces feature and prediction consistency between different image strips, b) DDM that effectively processes weak supervision signals, and c) DBC that handles strip boundary effects. → arrows indicate the flow for both training and testing, while → represent the flow only during training.

where the strip dimensions are determined by $h = H/n$, $w = W/m$ for horizontal strips, and $h = H/m$, $w = W/n$ for vertical strips. The optimal values of m and n are empirically determined through ablation studies in Section IV-F1. The image, vertical strips, and horizontal strips are separately processed in the main, horizontal, and vertical branches respectively, through a weight-shared U-Net backbone u, which is jointly optimized during training.

$$\mathbf{f}_h = \{u(\mathbf{x}_{i,j}) | \mathbf{x}_{i,j} \in \mathbf{x}_h\}, \tag{4}$$

$$\mathbf{f}_v = \{u(\mathbf{x}_{i,j}) | \mathbf{x}_{i,j} \in \mathbf{x}_v\}, \tag{5}$$

where $\mathbf{f}_h$ and $\mathbf{f}_v$ are sets of features $\mathbf{f}_{i,j} \in \mathbb{R}^{C' \times h \times w}$. These strip features are merged to reconstruct full-size feature maps,

$$\mathbf{F}^h = \mathcal{P}^{-1}(\mathbf{f}_h) \in \mathbb{R}^{C' \times H \times W}, \tag{6}$$

$$\mathbf{F}^v = \mathcal{P}^{-1}(\mathbf{f}_v) \in \mathbb{R}^{C' \times H \times W}, \tag{7}$$

where $\mathcal{P}^{-1}$ denotes the inverse operation of strip decomposition that restores the original spatial dimensions.

To ensure structural consistency and feature coherence between different views, we implement a hierarchical consistency learning strategy that operates at both feature and classification levels. At the feature level, we enforce strip feature consistency between corresponding strip features through Mean Squared Error,

$$\mathcal{L}_{sfc} = \|\mathbf{F}^h - \mathbf{F}^v\|_2^2, \tag{8}$$

where strip feature consistency ensures that the model learns view-invariant representations of road structures across different strip orientations.At the classification level, we enforce strip prediction consistency between corresponding strip predictions through Mean Absolute Error,

$$\mathcal{L}_{scc} = \|p^h_{aux} - p^v_{aux}\|_1, \tag{9}$$

where $\mathbf{p}^h_{aux}$ and $\mathbf{p}^v_{aux}$ are auxiliary predictions from DDM (Section III-B2). The Mean Absolute Error enforces prediction consistency between complementary strip views and maintains sharp road boundaries.

The final consistency loss combines both feature and classification consistency,

$$\mathcal{L}_{con} = \mathcal{L}_{sfc} + \mathcal{L}_{scc}. \tag{10}$$

This dual-level strip consistency regularization guides the model to learn robust road features while maintaining structural coherence across different strip views.

2) Dual Downsample Module(DDM): Traditional downsampling approaches in weakly supervised road segmentation often result in loss of fine-grained road features and structural discontinuities. To address this issue, we propose the DDM, which processes features through two parallel paths to better capture both global context and local details, as illustrated in Fig. 2 (b).

The DDM processes the input feature map $\mathbf{F}$ through two parallel paths: a standard path for global feature extraction and an enhanced path for fine-grained feature refinement. The standard path processes features through a segmentation head to obtain CAM features $\mathbf{F}_{CAM}$, which are then max-pooled to generate the main prediction,

$$\mathbf{p} = \text{MaxPool}(\mathbf{F}_{CAM}). \tag{11}$$

The enhanced path processes the input features through a learnable downsampling network E_{down} (based on ResNet34 [39]) to obtain features $\mathbf{F}_{down}$. These features are concatenated with the downsampled $\mathbf{F}_{CAM}$ to form $\mathbf{F}_{Concat}$. Subsequently, a Pixel Correlation Module (PCM) [20] computes the inter-pixel feature similarity on $\mathbf{F}_{Concat}$ via cosine distance,

$$f(\mathbf{x}_i, \mathbf{x}_j) = \frac{\text{conv}(\mathbf{x}_i)^T \text{conv}(\mathbf{x}_j)}{\|\text{conv}(\mathbf{x}_i)\| \cdot \|\text{conv}(\mathbf{x}_j)\|}, \tag{12}$$

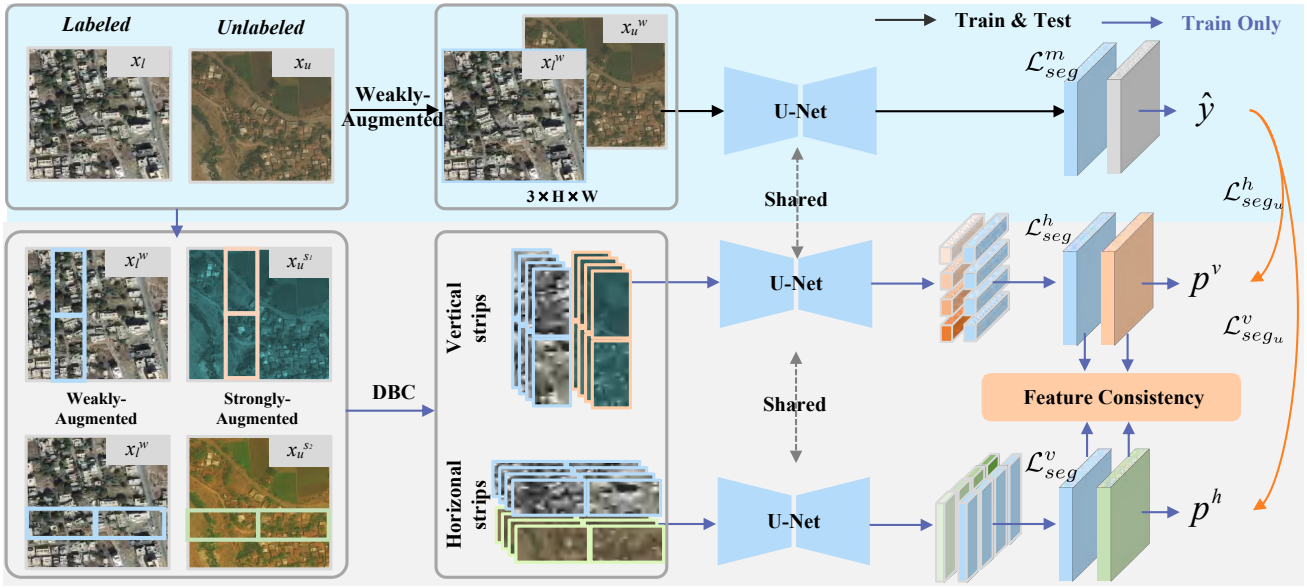

Fig. 3. Overview of our DualStrip-Net for semi-supervised road segmentation. The DualStrip-Net processes both labeled and unlabeled images through three parallel branches with shared U-Net backbone. The base branch (top) processes full images for supervised learning, while the horizontal and vertical strip branches (middle and bottom) enable consistency learning between different views. → arrows indicate the flow for both training and testing, while → represent the flow only during training.

where $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{F}_{Concat}$, and conv$(\cdot)$ represents the 1×1 convolution operation. The refined features are obtained through weighted aggregation,

$$\mathbf{f}_{att} = \frac{1}{C(\mathbf{x}_i)} \sum_{\forall j} \text{ReLU}(f(\mathbf{x}_i, \mathbf{x}_j))\mathbf{F}_{down,j}, \tag{13}$$

where $\mathbf{F}_{down,j}$ are the features from E_{down}, ReLU activation suppresses negative similarity values, and $C(\mathbf{x}_i)$ is the normalization factor. The auxiliary prediction is obtained by

$$\mathbf{p}_{aux} = \text{MaxPool}(\sigma(\text{cls}(\mathbf{f}_{att}) + \text{cls}(\mathbf{F}_{down}))), \tag{14}$$

where cls$(\cdot)$ denotes the segmentation head and σ is the sigmoid function. This dual-path design applies to all three branches (main, horizontal strip, and vertical strip), which enables comprehensive feature refinement across different views with efficient computation. The multi-perspective feature processing preserves both global road context and local structural details, which proves essential for road connectivity in weakly-supervised scenarios.

3) Dynamic Boundary-padded Cutting (DBC): The proposed DBC mechanism addresses the boundary effects in strip-based feature learning. This design is motivated by two observations in remote sensing imagery: 1) road structures exhibit spatial continuity across image divisions, and 2) road intersections contain complex geometric patterns that demand broader contextual information. When dividing an image into strips for processing, the road structures themselves naturally continue across these boundaries. Our DBC mechanism addresses the potential feature discontinuity through adaptive directional padding before cutting the image into strips.

$$\mathbf{x}_{pad} = \begin{cases} \mathcal{A}(\mathbf{x}, [\delta, \delta, 0, 0]), & \text{if } p = 0 \text{ (horizontal)}, \\ \mathcal{A}(\mathbf{x}, [0, 0, \delta, \delta]), & \text{if } p = 1 \text{ (vertical)}, \end{cases} \tag{15}$$

where $\mathbf{x}$ denotes the input image and $\mathcal{A}(\mathbf{x}, [t, b, l, r])$ denotes the adaptive padding operator that pads the image $\mathbf{x}$ with t, b, l, r pixels at the top, bottom, left, and right borders, respectively. For horizontal strips, padding is applied to the top and bottom borders, while for vertical strips, padding is applied to the left and right borders. This creates deliberate overlapping regions between adjacent strips, allowing the network to capture continuous road segments that would otherwise be divided at strip boundaries. During training, the padding size δ is dynamically selected from a predefined set $\Delta = \{0, p_1, ..., p_k\}$. Each value p_i is configured as an integer multiple of the encoder's downsampling factor (e.g., for an encoder with a downsampling factor of 32, $p_i = 32k_i$ where $k_i \in \mathbb{Z}^+$). The impact of different Δ configurations on model performance is thoroughly analyzed in Section IV-F4 of the experimental results.

The padded strips are processed through the shared backbone u as defined in Section III-B1. By carefully aligning the padding size δ, we ensure that the padded regions maintain feature consistency throughout the encoding-decoding process. This alignment effectively prevents feature misalignment and discontinuities at strip boundaries during reconstruction. When strip features are later merged to reconstruct the full-size feature map, road structures maintain their spatial continuity because each strip contains contextual information from its neighboring regions.

C. Loss Function

To effectively train our framework across different supervision settings, we design a unified loss function that consists of two complementary components: a segmentation loss that handles both weakly and fully supervised signals, and a consistency loss that enforces agreement between different

views. The overall loss function of our framework consists of two parts,

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{con}, \tag{16}$$

where for weakly-supervised setting

$$\mathcal{L}_{seg} = \sum_{i \in \{m,h,v\}} (\mathcal{L}^i_{seg} + \mathcal{L}^i_{aux}), \tag{17}$$

and for semi-supervised setting

$$\mathcal{L}_{seg} = \sum_{i \in \{m,h,v\}} \mathcal{L}^i_{seg} + \sum_{j \in \{h,v\}} \mathcal{L}^j_{seg_u}. \tag{18}$$

For all segmentation losses, we use cross-entropy as the loss function. Specifically, for the main and strip branches with labeled data:

$$\mathcal{L}^i_{seg} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=0}^{C-1} y_{n,c} \log(p^i_{n,c}), \tag{19}$$

where N is the number of pixels, C is the number of classes (typically 2 for road segmentation), $y_{n,c}$ is the ground truth label, and $p^i_{n,c}$ is the predicted probability from branch $i \in \{m, h, v\}$. Similarly, the auxiliary loss for the weakly-supervised setting is defined as:

$$\mathcal{L}^i_{aux} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=0}^{C-1} y_{n,c} \log(p^i_{aux,n,c}), \tag{20}$$

where $p^i_{aux,n,c}$ represents the auxiliary prediction from the DDM in branch i. For unlabeled data in the semi-supervised setting, we use pseudo-labels generated from the model's predictions:

$$\mathcal{L}^j_{seg_u} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=0}^{C-1} \hat{y}_{n,c} \log(p^j_{n,c}), \tag{21}$$

where $\hat{y}_{n,c}$ represents the pseudo-label generated from the model's confident predictions, and $j \in \{h, v\}$ indicates the horizontal or vertical strip branch. The parameter λ is empirically set to 1.

D. Overall Procedure for Road Segmentation

1) Weakly Supervised Road Segmentation: In the weakly-supervised setting, DualStrip-Net processes images through three parallel branches with a shared U-Net backbone, as illustrated in Fig. 2. The main branch processes full images to capture global context, while two strip branches (horizontal and vertical) exploit road characteristics through orthogonal directional views. Each branch incorporates three key components: 1) DualStrip Learning that enforces feature and prediction consistency between different image strips, 2) DDM that effectively processes weak supervision signals, and 3) DBC that handles strip boundary effects. This design enables effective learning from patch-level and scribble labels while maintaining road structural continuity.

2) Semi-Supervised Road Segmentation: The semi-supervised setting adapts DualStrip-Net to handle both labeled and unlabeled data simultaneously, as shown in Fig. 3. While maintaining the three-branch architecture and DBC component, the semi-supervised framework employs differential augmentation strategies: the base branch uses weak augmentations (e.g., random flip, random crop) for full images, while the strip branches adopt stronger augmentations that follow UniMatch [17] (ColorJitter, RandomGrayscale, and Gaussian blur). Unlike the weakly-supervised setting that employs both feature and prediction consistency, the semi-supervised framework only utilizes strip feature consistency ($\mathcal{L}_{sfc}$) between different views, which effectively regularizes the feature learning process while enabling efficient utilization of both labeled and unlabeled data for road segmentation.

IV. Experiments and Analyses

We evaluate the proposed DualStrip-Net for both weakly-supervised and semi-supervised road segmentation on four representative datasets with diverse road patterns. Through extensive experiments and ablation studies, we demonstrate state-of-the-art performance and analyze the effectiveness of each component.

A. Datasets

We introduce the datasets used in our experiments. To ensure comprehensive evaluation, we evaluate DualStrip-Net on four road segmentation benchmarks: DeepGlobe [40], Massachusetts Road [41], CHN6-CUG [42], and LSVG [26]. The first three datasets serve as standard benchmarks for road segmentation with pixel-wise annotations, while LSVG enables evaluation of cross-region generalization ability.

DeepGlobe Road [40] consists of 6,226 images, each with a resolution of 1024×1024 pixels and a ground sampling distance (GSD) of 0.5 m/pixel. The images are randomly divided into training and validation sets in a 3:1 ratio. To create non-overlapping crops of size 512×512 pixels, a total of 18,678 training samples and 6,226 validation samples are generated, as reported in previous studies [8] [9].

LSVG [26] includes images from several cities: Boston and its surrounding areas in the United States, Birmingham in the United Kingdom, and Shanghai in China. Unlike publicly available road datasets, the LSRV dataset features images with varying resolutions from different geographical regions, providing a comprehensive basis to test the generalizability of the model. Thus, the LSRV dataset is utilized to evaluate the generalization performance of weakly supervised road segmentation.

Massachusetts Road [41] comprises 1,171 aerial images from the state of Massachusetts. Each image, sized 1500×1500 pixels with a GSD of 1 m/pixel, is divided into three subsets: 1,108 images for training, 14 for validation, and 49 for testing. The dataset encompasses diverse urban, suburban, and rural areas, spanning over 2,600 square kilometers.

CHN6-CUG [42] The CHN6-CUG dataset, with a GSD of 0.5 m/pixel, contains 4,511 labeled images of size 512×512 pixels, partitioned into 3,608 images for training and 903 images for testing and evaluation.

B. Implementation Details and Experimental Setup

We implement DualStrip-Net using PyTorch and conduct all experiments on two NVIDIA RTX 3090 GPUs. We adopt the SGD optimizer with an initial learning rate of 0.01 and weight decay of 1×10^{-4}. The learning rate follows a polynomial decay schedule with power 2. A batch size of 8 is used consistently across all experiments.

For data augmentation, we employ different strategies in weakly-supervised and semi-supervised settings. In weakly-supervised experiments, we apply strong augmentations including random horizontal flipping, random rotation (0°, 90°, 180°, 270°), and random cropping (512×512) to enhance model performance. In semi-supervised experiments, to ensure fair comparison with existing methods, we only use basic augmentations (random horizontal flipping and random cropping) for the main branch. For dynamic boundary padding, we adaptively select padding sizes from 0, 32, 64, which helps maintain feature continuity at strip boundaries. We adopted U-Net [43] as our segmentation network with ResNet34 [39], which was pre-trained on ImageNet.

C. Evaluation Metrics

We employ five standard metrics to comprehensively evaluate the performance of road segmentation methods: Precision (P), Recall (R), F1-score (F1), Intersection over Union (IoU), and Boundary F1-score (BF1). Precision ($\mathrm{P} = \frac{TP}{TP+FP}$), which measures the proportion of correctly predicted road pixels among all pixels predicted as roads. Recall ($\mathrm{R} = \frac{TP}{TP+FN}$), which quantifies the proportion of actual road pixels correctly identified. F1-score ($\mathrm{F1} = \frac{2\times \mathrm{P}\times \mathrm{R}}{\mathrm{P}+\mathrm{R}}$), which provides a balanced measure between precision and recall. IoU ($\mathrm{IoU} = \frac{TP}{TP+FP+FN}$), which measures the overlap between the predicted road segmentation and the ground truth. BF1, which evaluates the accuracy of road boundary delineation following the implementation in [9] using MATLAB built-in functions. Higher values for all metrics indicate better performance. While IoU and F1-score evaluate overall segmentation quality, BF1 focuses on boundary accuracy, which is crucial for road network topology preservation.

D. Experiments in Weakly-supervised Road Segmentation

To comprehensively evaluate our proposed DualStrip-Net, we conduct extensive experiments on DeepGlobe Road [40] and LSVG datasets [26]. The experimental analysis compares the performance with existing weakly-supervised methods and evaluates the segmentation ability in diverse road scenarios.

1) Analysis on DeepGlobe Dataset: As shown in Table I, DualStrip-Net-$\mathcal{P}_{64}$ achieves state-of-the-art performance with 77.09% F1-score and 62.72% IoU on the DeepGlobe dataset. Compared to the baseline UNet-CAM which also uses patch-level labels, our approach (DualStrip-Net-$\mathcal{P}_{64}$) shows substantial improvements (+3.33% F1-score, +4.29% IoU). This significant gain demonstrates the effectiveness of our DualStrip Learning mechanism in better exploiting patch-level supervision. Our method also outperforms ScRoadExtractor (74.47% F1-score) which uses scribble-based supervision. The high Boundary F1-score of 84.42% (+2.50% over SOC-RoadNet)

TABLE I
PERFORMANCE COMPARISON OF ROAD SEGMENTATION METHODS ON THE DEEPGLOBE DATASET. † DENOTES RESULTS REPORTED IN ORIGINAL PAPERS. $\mathcal{F}$ DENOTES THE FULLY-SUPERVISED BASELINE, $\mathcal{P}$ DENOTES PATCH-LEVEL SUPERVISION, AND $\mathcal{S}$ DENOTES SCRIBBLE-LEVEL SUPERVISION. **BOLD** REPRESENTS OUR PROPOSED DUALSTRIP-NET. MAGENTA AND CYAN DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Methods	Sup.	P	R	F1	IoU	BF1
U-Net [43]	$\mathcal{F}$	83.76	75.39	79.35	65.77	82.64
LinkNet [44]	$\mathcal{F}$	81.98	76.60	79.20	65.56	86.33
D-LinkNet [45]	$\mathcal{F}$	80.43	79.01	79.71	66.27	87.04
DeepLabv3+ [46]	$\mathcal{F}$	80.81	77.51	79.13	65.46	85.09
GAMSNet† [26]	$\mathcal{F}$	83.14	80.00	81.54	68.84	-
ScribbleSaliency [10]	$\mathcal{S}$	51.01	67.79	58.21	41.06	55.97
WSL4MIS [47]	$\mathcal{S}$	51.19	84.69	63.81	46.86	65.32
Scribble2label [48]	$\mathcal{S}$	61.74	78.39	69.08	52.76	73.50
ScRoadExtractor [8]	$\mathcal{S}$	69.39	80.35	74.47	59.33	81.89
SOC-RoadNet† [9]	$\mathcal{S}$	77.28	73.16	-	60.21	81.92
DualStrip-Net-$\mathcal{S}$(ours)	$\mathcal{S}$	**69.74**	**81.08**	**74.98**	**59.98**	**82.35**
UNet-CAM [13](baseline)	$\mathcal{P}$	69.66	78.38	73.76	58.43	82.47
DualStrip-Net-$\mathcal{P}_{64}$(ours)	$\mathcal{P}$	**76.49**	**77.70**	77.09	62.72	84.42
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**77.41**	**80.31**	78.84	65.06	86.05

further validates the segmentation ability in preserving structural details. The experiments reveal a critical relationship between patch size and segmentation performance. Pixel-level labels (1×1 patches) that provide the most detailed supervision require prohibitive annotation costs. Our method addresses this challenge through patch-level labels. The reduction in patch size leads to improved segmentation performance due to more precise boundary annotations. The results indicate that an appropriate patch size achieves an effective balance between annotation precision and annotation cost.

The visual results in Fig. 4 demonstrate the superior performance of DualStrip-Net in various challenging scenarios. In complex intersection areas (Fig. 4(a)), our method successfully maintains road connectivity while accurately capturing the intersection topology. For parallel roads (Fig. 4(d)), DualStrip-Net effectively distinguishes individual roads and maintains proper spacing, avoiding the common issue of road merging. The effectiveness of our feature learning is visually evident in Fig. 9. While UNet-CAM produces fragmented predictions from patch labels, our DualStrip mechanism generates more focused and coherent activation maps. This improvement in feature discrimination is particularly noticeable in areas with road-like textures (e.g., building boundaries), where our method successfully reduces false positives. However, our visual analysis also reveals certain limitations. In dense buildings or heavy shadows areas (Fig. 4), the method occasionally struggles to maintain consistent road detection. These challenging cases suggest potential directions for future improvements, particularly in handling complex urban environments with varying illumination conditions.

2) Cross-region Generalization Analysis: To evaluate the generalization ability of our method, we conduct extensive experiments on three diverse urban regions: Boston, Birmingham, and Shanghai. Specifically, Tables II, III, and IV present the strong generalization ability of DualStrip-Net

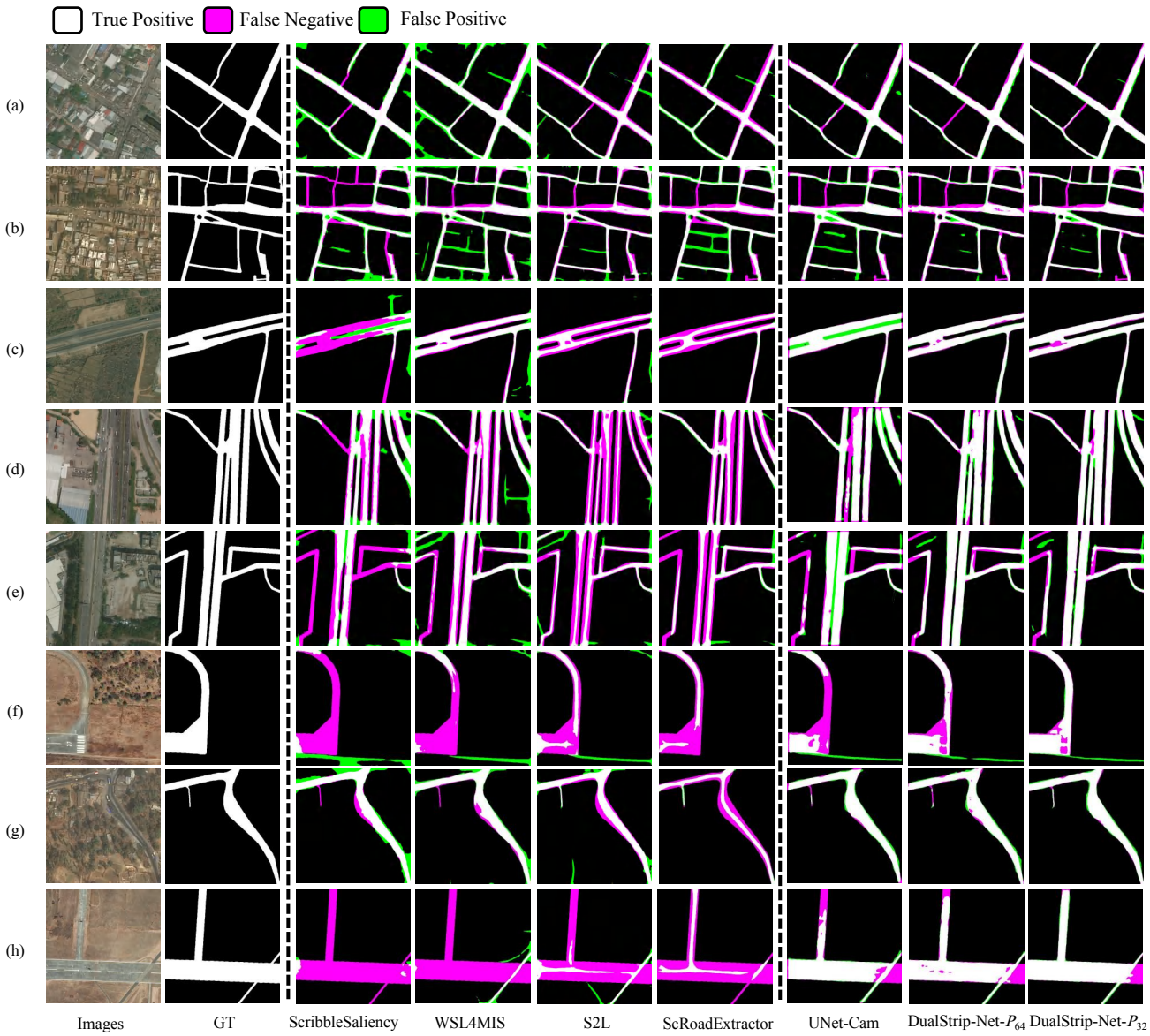

Fig. 4. Qualitative results of our proposed DualStrip-Net and other baseline methods for the comparison on DeepGlobe dataset.

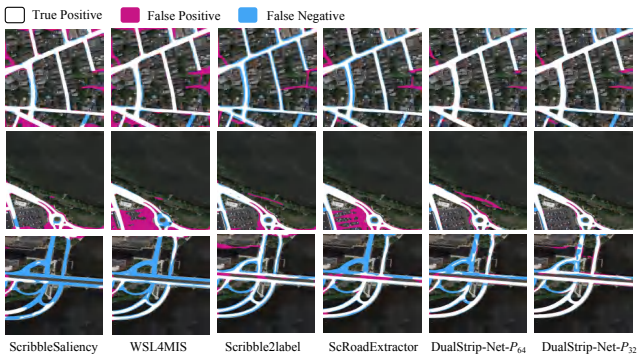

Fig. 5. Cross-region generalization results on Boston dataset.

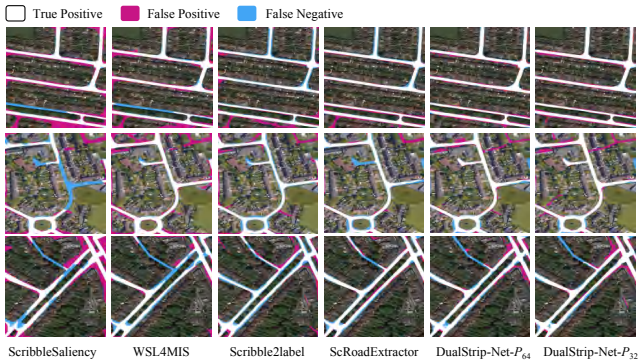

Fig. 6. Cross-region generalization results on Birmingham dataset.

across these different regions. Using the same patch size of 64×64, our DualStrip-Net-$\mathcal{P}_{64}$ consistently outperforms the baseline UNet-CAM with F1 score improvements of +0.34%, +0.72%, and +3.73% on Boston, Birmingham, and Shanghai respectively. In particular, the performance stability across these diverse regions stands out significantly, with IoU improvements of +0.43%, +0.83%, and +3.95% over UNet-CAM. Each city presents unique characteristics, where

TABLE II
QUANTITATIVE COMPARISON OF ROAD SEGMENTATION METHODS ON BOSTON DATASET.

Methods	Sup.	P	R	F1	IoU
U-Net [43]	$\mathcal{F}$	86.83	67.96	76.24	61.61
LinkNet [44]	$\mathcal{F}$	85.41	69.93	76.90	62.46
D-LinkNet [45]	$\mathcal{F}$	82.50	75.27	78.72	64.90
DeepLabv3Plus [46]	$\mathcal{F}$	85.45	70.61	77.32	63.03
GAMSNet† [26]	$\mathcal{F}$	87.46	69.36	77.37	63.09
ScribbleSaliency [10]	$\mathcal{S}$	69.36	64.96	67.09	50.47
WSL4MIS [47]	$\mathcal{S}$	68.78	69.07	68.92	52.58
scribble2label [48]	$\mathcal{S}$	78.68	58.01	66.78	50.13
ScRoadExtractor [8]	$\mathcal{S}$	83.73	67.59	74.80	59.75
DualStrip-Net-$\mathcal{S}$(ours)	$\mathcal{S}$	**85.25**	**66.26**	**74.57**	**59.45**
UNet-CAM [13] (Baseline)	$\mathcal{P}$	78.64	70.14	74.15	58.92
DualStrip-Net-$\mathcal{P}_{64}$(ours)	$\mathcal{P}$	**83.01**	**67.55**	**74.49**	**59.35**
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**84.80**	**69.62**	**76.46**	**61.90**

TABLE III
QUANTITATIVE COMPARISON OF ROAD SEGMENTATION METHODS ON BIRMINGHAM DATASET.

Methods	Sup.	P	R	F1	IoU
U-Net [43]	$\mathcal{F}$	72.68	74.57	73.61	58.25
LinkNet [44]	$\mathcal{F}$	70.26	75.54	72.80	57.23
D-LinkNet [45]	$\mathcal{F}$	68.62	78.97	73.44	58.02
DeepLabv3Plus [46]	$\mathcal{F}$	65.30	81.63	72.56	56.93
GAMS-Net† [26]	$\mathcal{F}$	74.77	70.13	72.38	56.71
ScribbleSaliency [10]	$\mathcal{S}$	54.46	64.08	58.88	41.72
WSL4MIS [47]	$\mathcal{S}$	52.15	76.48	62.01	44.95
Scribble2label [48]	$\mathcal{S}$	65.00	66.27	65.63	48.84
ScRoadExtractor [8]	$\mathcal{S}$	71.33	71.58	**71.45**	**55.59**
DualStrip-Net-$\mathcal{S}$(ours)	$\mathcal{S}$	**75.16**	**65.60**	70.05	53.91
UNet-CAM [13] (Baseline)	$\mathcal{P}$	60.92	77.90	68.37	51.94
DualStrip-Net-$\mathcal{P}_{64}$(ours)	$\mathcal{P}$	**66.26**	**72.17**	**69.09**	**52.77**
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**61.16**	**78.60**	**68.79**	**52.43**

TABLE IV
QUANTITATIVE COMPARISON OF ROAD SEGMENTATION METHODS ON SHANGHAI DATASET.

Methods	Sup.	P	R	F1	IoU
U-Net [43]	$\mathcal{F}$	83.55	49.15	61.89	44.81
LinkNet [44]	$\mathcal{F}$	80.02	53.00	63.76	46.80
D-LinkNet [45]	$\mathcal{F}$	84.02	51.16	63.60	46.63
DeepLabv3Plus [46]	$\mathcal{F}$	52.33	83.05	64.21	47.28
GAMSNet$\mathcal{S}$† [26]	$\mathcal{F}$	87.70	48.31	62.30	45.25
ScribbleSaliency [10]	$\mathcal{S}$	64.38	40.03	49.36	32.77
WSL4MIS [47]	$\mathcal{S}$	57.92	49.36	53.30	36.33
Scribble2label [48]	$\mathcal{S}$	74.74	42.14	53.90	36.89
ScRoadExtractor [8]	$\mathcal{S}$	83.76	46.81	60.05	42.91
DualStrip-Net-$\mathcal{S}$(ours)	$\mathcal{S}$	**82.45**	**46.00**	**59.05**	**41.90**
UNet-CAM [13] (Baseline)	$\mathcal{P}$	74.95	50.65	60.45	43.31
DualStrip-Net-$\mathcal{P}_{64}$(ours)	$\mathcal{P}$	**79.04**	**54.03**	64.18	47.26
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**80.48**	**56.03**	**66.07**	**49.33**

TABLE V
COMPREHENSIVE EVALUATION OF SEMI-SUPERVISED ROAD SEGMENTATION METHODS ON THE DEEPGLOBE ROAD DATASET UNDER DIFFERENT LABELED DATA RATIOS (5%, 10%, AND 20%). † DENOTES RESULTS CITED FROM [19]. **BOLD** REPRESENTS OUR PROPOSED DUALSTRIP-NET. MAGENTA AND CYAN DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Method	Sup.	5%		10%		20%	
		F1	IoU	F1	IoU	F1	IoU
s4GAN [49]	$\mathcal{F}$	59.83	42.68	62.76	45.73	63.53	46.55
ST++ [15]	$\mathcal{F}$	67.72	51.20	70.64	54.60	73.23	57.77
FixMatch [16]	$\mathcal{F}$	72.16	56.44	74.59	59.47	75.96	61.24
UniMatch [17]	$\mathcal{F}$	72.78	57.21	74.12	58.88	76.30	61.68
AdaptMatch† [50]	$\mathcal{F}$	58.70	41.50	59.90	42.80	63.40	46.40
SemiRoadExNet [18]	$\mathcal{F}$	67.98	51.50	70.07	53.92	73.50	58.11
MCMCNet† [19]	$\mathcal{F}$	68.20	51.80	71.20	55.40	73.80	58.40
Full-supervised (79.35/65.77)	$\mathcal{F}$	66.33	49.62	67.57	51.02	73.13	57.65
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**68.69**	**52.31**	**71.08**	**55.14**	**72.40**	**56.75**
DualStrip-Net(ours)	$\mathcal{F}$	74.75	59.69	75.68	60.87	76.98	62.57

Boston features organized grid systems, Birmingham contains historical road layouts, and Shanghai exhibits dense urban infrastructure. Notably, the DualStrip mechanism captures generalizable road features through complementary views and reduces the risk of overfitting to region-specific patterns.

The visual results in Figures 5, 6, and 7 demonstrate the adaptability of our method to diverse urban scenarios. In Boston (Fig. 5), our method successfully handles the structured road networks with clear true positive predictions shown in white. Similarly, in Birmingham (Fig. 6), DualStrip-Net effectively adapts to the varying road widths and complex urban patterns, with minimal false positives marked in blue. Moreover, the method shows robust performance in Shanghai (Fig. 7), where dense building clusters create challenging road environments. The qualitative comparison uses a consistent color scheme across all regions, where white indicates true positive predictions, magenta shows true negatives, and blue represents false positives. This visualization clearly demonstrates our method ability to maintain high precision across different urban contexts. Finally, the analysis reveals region specific challenges through these visualizations, while validating the strong cross region generalization ability of our approach in complex urban environments.

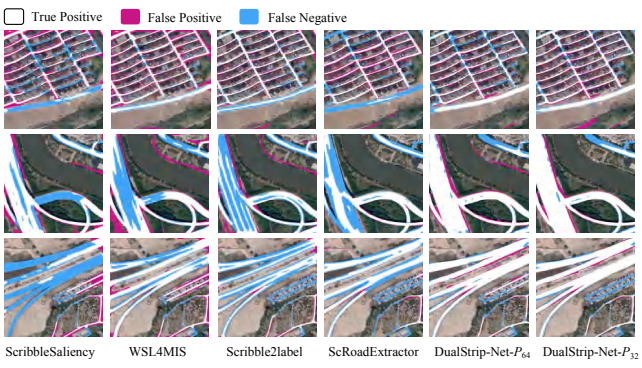

Fig. 7. Cross-region generalization results on Shanghai dataset.

E. Experiments in Semi-supervised Road Segmentation

1) Results on DeepGlobe Dataset: As shown in Table V, DualStrip-Net demonstrates consistent performance gains across different label ratios on DeepGlobe dataset, achieving F1/IoU scores from 74.75%/59.69% (5% labels) to 76.98%/62.57% (20% labels). This scalability out-

TABLE VI
COMPREHENSIVE EVALUATION OF SEMI-SUPERVISED ROAD SEGMENTATION METHODS ON THE MASSACHUSETTS DATASET UNDER DIFFERENT LABELED DATA RATIOS (5%, 10%, AND 20%).

Method	Sup.	5%		10%		20%	
		F1	IoU	F1	IoU	F1	IoU
s4GAN [49]	$\mathcal{F}$	65.41	48.60	67.35	50.77	68.16	51.70
ST++ [15]	$\mathcal{F}$	67.16	50.56	68.09	51.62	69.00	52.67
FixMatch [16]	$\mathcal{F}$	73.61	58.24	73.38	57.96	73.75	58.41
UniMatch [17]	$\mathcal{F}$	73.95	58.67	73.64	58.28	74.38	59.21
AdaptMatch† [50]	$\mathcal{F}$	56.80	39.70	57.40	40.20	57.60	40.50
SemiRoadExNet [18]	$\mathcal{F}$	68.54	52.14	72.68	57.08	73.48	58.07
MCMCNet† [19]	$\mathcal{F}$	72.20	56.50	73.40	58.00	74.60	59.50
Full-supervised (75.36/60.47)	$\mathcal{F}$	59.39	42.24	65.41	48.60	70.94	54.96
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**66.17**	**49.44**	**69.04**	**52.72**	**70.87**	**54.88**
DualStrip-Net(ours)	$\mathcal{F}$	**75.25**	**60.32**	**75.11**	**60.15**	**76.33**	**61.73**

TABLE VII
COMPREHENSIVE EVALUATION OF SEMI-SUPERVISED ROAD SEGMENTATION METHODS ON THE CHN6-CUG DATASET UNDER DIFFERENT LABELED DATA RATIOS (5%, 10%, AND 20%).

Method	Sup.	5%		10%		20%	
		F1	IoU	F1	IoU	F1	IoU
s4GAN [49]	$\mathcal{F}$	51.73	34.89	56.28	39.16	61.29	44.18
ST++† [15]	$\mathcal{F}$	22.30	12.60	31.40	18.60	37.30	23.00
FixMatch [16]	$\mathcal{F}$	69.94	53.78	73.02	57.50	72.82	57.25
UniMatch [17]	$\mathcal{F}$	71.12	55.18	71.06	55.11	71.88	56.11
AdaptMatch† [50]	$\mathcal{F}$	39.00	24.20	58.60	41.50	65.90	48.30
SemiRoadExNet† [18]	$\mathcal{F}$	61.80	43.44	62.85	44.88	66.83	48.67
MCMCNet† [19]	$\mathcal{F}$	65.70	48.90	68.90	52.60	71.30	55.70
Full-supervised (73.58/58.20)	$\mathcal{F}$	56.23	39.11	64.51	47.61	67.96	51.47
DualStrip-Net-$\mathcal{P}_{32}$(ours)	$\mathcal{P}$	**62.50**	**45.45**	**67.64**	**51.11**	**70.20**	**54.08**
DualStrip-Net(ours)	$\mathcal{F}$	**71.00**	**55.04**	**72.86**	**57.30**	**73.91**	**58.61**

performs both consistency-based methods like UniMatch (72.78%/57.21% to 76.30%/61.68%) and road-specific methods like MCMCNet (68.20%/51.80% to 73.80%/58.40%). The model with patch labels ($\mathcal{P}_{32}$) shows competitive performance, which validates the effectiveness of strip-based consistency regularization in compensating for coarse supervision through bi-directional feature alignment. As illustrated in Fig. 8 (bottom row), DualStrip-Net effectively handles rural road scenarios. The model accurately segments the road network that connects scattered residential areas and farmland, maintaining road continuity despite the sparse context. The clear segmentation boundaries between roads and the surrounding rural terrain demonstrate the robustness in areas with limited structural cues.

2) Results on Massachusetts Dataset: The experiments on Massachusetts dataset reveal a significant finding in Table VI: with only 20% labeled data, DualStrip-Net (76.33% F1-score, 61.73% IoU) surpasses the fully-supervised baseline (75.36% F1-score, 60.47% IoU). This result demonstrates the effectiveness of representation learning from unlabeled data: while labeled samples provide direct supervision signals, the strip-based consistency regularization on unlabeled data enhances feature discrimination and structural awareness. When using patch labels ($\mathcal{P}_{32}$), the model maintains strong performance. The strip-based approach addresses the limited receptive field issue through multi-directional context aggregation, which enables effective representation learning despite the coarse patch

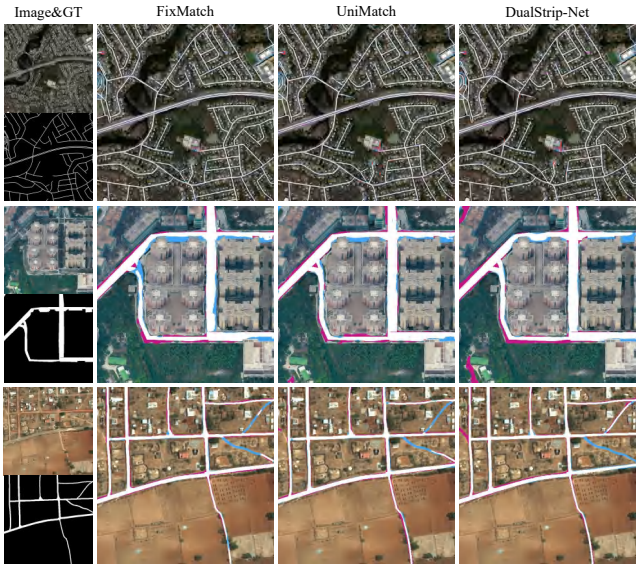

Fig. 8. Visual comparison of semi-supervised road segmentation with 20% labeled data. From top to bottom: Massachusetts, CHN6-CUG, and DeepGlobe datasets.

supervision. The visual results in Fig. 8 (top row) demonstrate the effectiveness in complex urban environments. DualStrip-Net accurately segments the curved roads and preserves the connectivity at multiple road intersections, where the road network exhibits intricate branching patterns. The clear road delineation against the dense residential areas validates the ability to handle urban complexity.

3) Results on CHN6-CUG Dataset: As demonstrated in Table VII, DualStrip-Net achieves consistent performance gains on CHN6 dataset with limited supervision, showing steady improvements from 5% to 20% label ratios. The significant performance boost from 5% to 10% labeled data validates the effectiveness of dualstrip learning. With patch supervision ($\mathcal{P}_{32}$), the model maintains robust performance across different label ratios. The effectiveness on wider roads stems from two aspects: 1) wider roads that provide larger receptive fields for feature consistency learning, which facilitates effective representation learning from unlabeled data; 2) the asymmetric augmentation strategy between main and strip branches that introduces diverse perturbations, which enhances the invariance to appearance variations. As shown in Fig. 8 (middle row), DualStrip-Net achieves precise segmentation of wide roads in industrial areas. The model accurately captures the sharp turns and maintains consistent road boundaries, where the road width varies significantly.

These comprehensive experiments demonstrate that DualStrip-Net effectively minimizes the performance gap between limited supervision and full supervision through strip-based representation learning. The effectiveness can be attributed to: 1) the complementary optimization between supervised learning and consistency regularization, which works well for both pixel-level and patch-level supervision, 2) the structural prior encoded through directional feature alignment, and 3) the adaptive receptive field modulation through dynamic boundary padding.

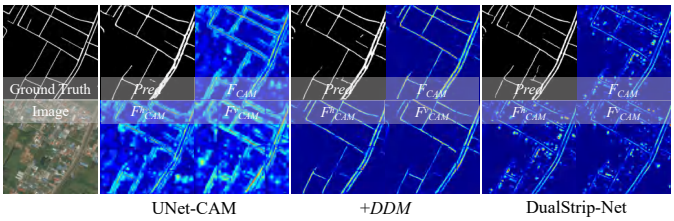

Fig. 9. Visualization of CAM features from different model variants on the DeepGlobe dataset.

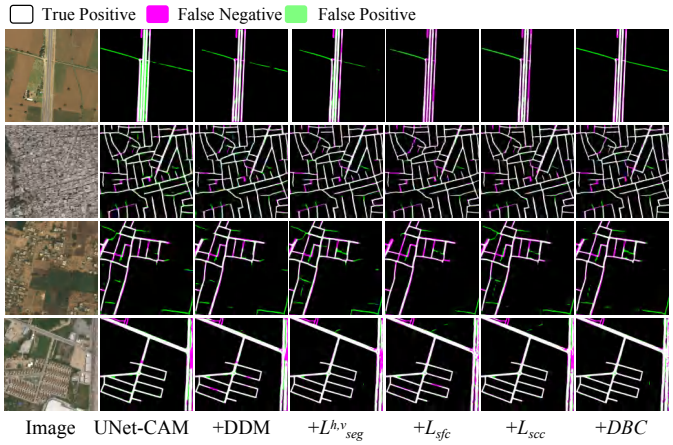

Fig. 10. Progressive improvements from different components on the DeepGlobe dataset. From left to right: input image, predictions from UNet-CAM (baseline), +DDM, $+\mathcal{L}_{seg}^{h,v}$, $+\mathcal{L}_{sfc}$, $+\mathcal{L}_{scc}$, and +DBC.

F. Ablation Studies

1) Different Strip Numbers under Weakly-supervised Setting: To investigate the impact of strip numbers on model performance, we conduct experiments with different strip configurations on the DeepGlobe dataset under weakly-supervised setting (DualStrip-Net-$\mathcal{P}_{64}$). As shown in Table VIII, we evaluate three configurations: 4×1, 4×2, and 8×2. The 4×2 configuration achieves the best performance with 62.72% IoU and 77.09% F1 score. The performance differences among these configurations can be attributed to several factors. First, compared to 4×1, the 4×2 configuration introduces complementary vertical strips that enhance road connectivity patterns, which leads to a 0.31% IoU improvement. This result validates our assumption about the complementary structural information from orthogonal strip views. However, when the configuration increases to 8×2, the performance drops by 1.06% IoU, which suggests adverse effects from excessive strip division. This degradation occurs due to three main reasons: 1) small strips that lack sufficient context for reliable road feature extraction, especially at complex intersections; 2) additional strip boundaries that create feature matching challenges; and 3) increased computational overhead that affects the stability of feature consistency. These results demonstrate that the optimal strip configuration depends on the trade-off between feature granularity and structural completeness. The 4×2 configuration achieves a balance with sufficient context in each strip and effective cross-strip feature correspondence.

2) Component Analysis under Weakly-supervised Setting: As shown in Table IX, we analyze the contribution of each proposed component. The DDM improves the baseline by capturing multi-scale road features through parallel downsampling paths, effectively handling roads of varying widths. While PuzzleCAM [22] introduces patch-based regularization for general object localization, our DualStrip mechanism is specifically designed for road segmentation. As shown in our ablation study (Table IX), PuzzleCAM achieves only marginal improvements (+0.87% F1, +1.10% IoU) over the baseline, while our DualStrip Learning brings substantial gains (+2.20% F1, +2.81% IoU). This performance gap demonstrates that explicitly modeling road structures through directional strips is more effective than generic patch-based learning for road segmentation tasks. The effectiveness of DualStrip is further validated in Table X, where the segmentation loss ($\mathcal{L}_{seg}$) from CAM-based pseudo labels on both main branch and strip branches provides the foundation for weakly-supervised learning. The feature consistency ($\mathcal{L}_{sfc}$) and classification consistency ($\mathcal{L}_{scc}$) then progressively enhance road feature learning. Finally, the DBC module further improves performance by introducing diverse boundary conditions during training, leading to total gains of 3.33% F1-score and 4.29% IoU over the baseline.

3) Qualitative Analysis under Weakly-supervised Setting: The qualitative analysis in Fig. 10 demonstrates how each component contributes to addressing specific challenges in road segmentation. DDM significantly mitigates false positives in areas with road-like textures by exploiting multi-scale context. The strip segmentation loss ($\mathcal{L}_{seg}^{h,v}$) enhances road continuity through learning from complementary orthogonal perspectives, while strip feature consistency ($\mathcal{L}_{sfc}$) helps maintain uniform road width by aligning features across different views. The strip classification consistency ($\mathcal{L}_{scc}$) further enhances structural coherence by enforcing consistent class predictions between complementary strips. The DBC improves robustness by training the model with varying boundary conditions, particularly effective in handling complex road boundaries and intersections.

TABLE VIII
ABLATION STUDY ON STRIP NUMBERS ON DEEPGLOBE DATASET UNDER WEAKLY-SUPERVISED SETTING (DUALSTRIP-NET-$\mathcal{P}_{64}$).

Strip Config	Metrics	
$m \times n$ (horizontal) / $n \times m$ (vertical)	IoU	F1
4×1 / 1×4	62.41	76.86
4×2 / 2×4	62.72	77.09
8×2 / 2×8	61.66	76.28

TABLE IX
ABLATION STUDY OF KEY COMPONENTS IN DUALSTRIP-NET ON DEEPGLOBE DATASET.

DDM	DualStrip	PuzzleCAM [22]	DBC	F1	IoU
				73.76	58.43
✓				74.80 (+1.04)	59.74 (+1.31)
		✓		74.63 (+0.87)	59.53 (+1.10)
	✓			75.96 (+2.20)	61.24 (+2.81)
✓	✓			76.69 (+2.93)	62.20 (+3.77)
✓	✓		✓	77.09 (+3.33)	62.72 (+4.29)

TABLE X
COMPONENT ANALYSIS OF DUALSTRIP LEARNING ON DEEPGLOBE DATASET UNDER WEAKLY-SUPERVISED SETTING(DUALSTRIP-NET-$\mathcal{P}$32).

DualStrip Learning			F1	IoU
$\mathcal{L}_{seg}$	$\mathcal{L}_{sfc}$	$\mathcal{L}_{scc}$		
✓			75.66	60.84
✓	✓		76.51	61.95
✓		✓	75.83	61.07
✓	✓	✓	76.69	62.20

TABLE XI
ABLATION STUDY ON DIFFERENT Δ VALUES FOR PATCH-LEVEL SUPERVISION UNDER 20% LABELED DATA SETTING

Δ	CHN6-CUG		Massachusetts	
	F1	IoU	F1	IoU
{0}	67.80	51.29	70.54	54.42
{0, 32}	68.52	52.12	70.76	54.75
{0, 32, 64}	70.20	54.08	70.87	54.88

4) Parameter and Component Analysis under Semi-supervised Setting: We further investigated the impact of different Δ values on model performance, as shown in Table XI. For the CHN6-CUG and Massachusetts dataset, as Δ values increase (from a single value of {0} to including {0, 32}, and then to {0, 32, 64}), the F1 and IoU metrics show a clear upward trend. This indicates that DBC can capture richer feature information, effectively enhancing road extraction accuracy.

For the Massachusetts dataset, we observed a slight decrease in F1 score (from 74.75% to 70.87%), while the IoU metric steadily improved (from 54.42% to 54.88%). This phenomenon may be attributed to differences in road feature distributions between the two datasets, where multi-scale δ values provide more comprehensive feature representations in complex scenes, although they might introduce redundant information in some simpler scenarios. The semi-supervised experiments (Table XII) reveal distinct behavior patterns across different datasets and labeling ratios. With only 5% labeled data, the baseline model struggles on both datasets (Massachusetts: 42.24% IoU, CHN6: 39.11% IoU). The segmentation loss ($\mathcal{L}_{seg}$) brings substantial improvements through complementary views. However, the impact of consistency mechanisms varies significantly with road characteristics. On Massachusetts dataset with narrow roads (7 pixels [41]), the performance stabilizes after incorporating $\mathcal{L}_{seg}$, whereas on CHN6-CUG with wider roads, the consistency regularization demonstrates consistent and steady improvements across different label ratios. This difference reveals that the effectiveness of consistency learning depends critically on the available spatial context(narrow roads provide insufficient feature matching cues), while wider roads enable more effective utilization of consistency constraints. These insights suggest that the optimal semi-supervised strategy should consider both the labeling ratio and the physical characteristics of roads in the target dataset.

TABLE XII
ABLATION STUDY ON MASSACHUSETTS AND CHN6-CUG DATASETS UNDER SEMI-SUPERVISED SETTING.

Ratio	DualStrip		DBC	Massachusetts		CHN6-CUG	
	$\mathcal{L}_{seg}$	$\mathcal{L}_{sfc}$		IoU	F1	IoU	F1
5%				42.24	59.37	39.11	56.24
	✓			60.31	75.24	53.59	69.78
	✓	✓		60.40	75.31	54.02	70.15
	✓		✓	60.72	75.56	53.80	69.96
	✓	✓	✓	60.32	75.25	55.04	71.00
10%				48.60	65.41	47.61	64.51
	✓			60.73	75.57	57.41	72.94
	✓	✓		60.54	75.42	56.57	72.26
	✓		✓	60.77	75.60	57.37	72.91
	✓	✓	✓	60.15	75.11	57.30	72.86
20%				54.96	70.94	51.47	67.94
	✓			61.96	76.52	58.34	73.69
	✓	✓		61.40	76.08	58.21	73.59
	✓		✓	61.72	76.33	58.05	73.46
	✓	✓	✓	61.73	76.33	58.61	73.91

V. CONCLUSION

In this work, we present DualStrip-Net, a novel framework that unifies strip-based learning and multi-scale feature extraction for weakly- and semi-supervised road segmentation. The framework introduces two key technical contributions: 1) a dualstrip learning mechanism that captures road structures from complementary perspectives, and 2) a Dual Downsampling Module that enhances multi-scale feature learning. Through comprehensive experiments on three challenging datasets (DeepGlobe, Massachusetts, and CHN6-CUG), we establish new benchmarks for road segmentation with limited annotations. The unified framework integrates strip-based segmentation with consistency learning, while the DBC enhances model robustness through diverse boundary conditions. The experimental results demonstrate that DualStrip-Net significantly outperforms existing methods in both weakly-supervised and semi-supervised settings.

REFERENCES

[1] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "Earthnets: Empowering artificial intelligence for earth observation," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–36, 2024.

[2] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–10, 2024.

[3] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multi-branch mutual-guiding learning for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[4] L. Ding and L. Bruzzone, "Diresnet: Direction-aware residual network for road extraction in vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10 243–10 254, 2021.

[5] Z. Luo, K. Zhou, Y. Tan, X. Wang, R. Zhu, and L. Zhang, "Ad-roadnet: An auxiliary-decoding road extraction network improving connectivity while preserving multiscale road details," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8049–8062, 2023.

[6] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vision. (ICCV)*, 2017, pp. 3458–3466.

[7] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2016, pp. 3159–3167.

[8] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[9] M. Zhou, H. Sui, S. Chen, J. Liu, W. Shi, and X. Chen, "Large-scale road extraction from high-resolution remote sensing images based on a weakly-supervised structural and orientational consistency constraint network," *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 234–251, 2022.

[10] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2020, pp. 12 543–12 552.

[11] R. Lian and L. Huang, "Weakly supervised road segmentation in high-resolution remote sensing images using point annotations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, 2016, pp. 549–565.

[13] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sens.*, vol. 12, no. 2, 2020.

[14] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2021, pp. 2613–2622.

[15] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2022.

[16] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[17] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2023.

[18] H. Chen, Z. Li, J. Wu, W. Xiong, and C. Du, "Semiroadexnet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 169–183, 2023.

[19] L. Gao, Y. Zhou, J. Tian, W. Cai, and Z. Lv, "Mcmcnet: A semi-supervised road extraction network for high-resolution remote sensing images via multiple consistency and multitask constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.

[20] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2020, pp. 12 272–12 281.

[21] J. Hu, J. Gao, Y. Yuan, J. Chanussot, and Q. Wang, "Lgnet: Location-guided network for road extraction from satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[22] S. Jo and I.-J. Yu, "Puzzle-cam: Improved localization via matching partial and full features," in *IEEE Int. Conf. Inf. Process.* IEEE, 2021, pp. 639–643.

[23] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully convolutional network-based ensemble method for road extraction from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, pp. 1777–1781, 2020.

[24] T. Sun, Z. Chen, W. ji Yang, and Y. Wang, "Stacked u-nets with multi-output for road extraction," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops(CVPRW)*, pp. 187–1874, 2018.

[25] Z. Yang, W. Zhang, Q. Li, W. Ni, J. Wu, and Q. Wang, "C²net: Road extraction via context perception and cross spatial-scale feature interaction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[26] X. Lu, Y. Zhong, Z. Zheng, and L. Zhang, "Gamsnet: Globally aware road detection network with multi-scale residual learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 340–352, 2021.

[27] G. Yuan, J. Li, X. Liu, and Z. Yang, "Weakly supervised road network extraction for remote sensing image based scribble annotation and adversarial learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 7184–7199, 2022.

[28] J. Feng, H. Huang, J. Zhang, W. Dong, D. Zhang, and L. Jiao, "Samixnet: Structure-aware mixup and invariance learning for scribble-supervised road extraction in remote sensing images," *ArXiv*, vol. abs/2403.01381, 2024.

[29] S. Wu, C. Du, H. Chen, Y. Xu, N. Guo, and N. Jing, "Road extraction from very high resolution images using weakly labeled openstreetmap centerline," *ISPRS Int. J. Geo Inf.*, vol. 8, p. 478, 2019.

[30] H. Chen, S. Peng, C. Du, J. Li, and S. Wu, "Sw-gan: Road extraction from remote sensing imagery using semi-weakly supervised adversarial learning," *Remote Sens.*, vol. 14, p. 4145, 2022.

[31] S. Kho, P. Lee, W. Lee, M. Ki, and H. Byun, "Exploiting shape cues for weakly supervised semantic segmentation," *Pattern Recognit.*, vol. 132, p. 108953, 2022.

[32] H. Zhang, P. Li, X. Liu, X. Yang, and L. An, "An iterative semi-supervised approach with pixel-wise contrastive loss for road extraction in aerial images," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, pp. 1–21, 2023.

[33] Z.-X. Yang, Z.-H. You, S.-B. Chen, J. Tang, and B. Luo, "Semisupervised edge-aware road extraction via cross teaching between cnn and transformer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8353–8362, 2023.

[34] X. Gu, S. Yu, F. Huang, S. Ren, and C. Fan, "Consistency self-training semi-supervised method for road extraction from remote sensing images," *Remote Sens.*, vol. 16, no. 21, 2024.

[35] J. Wang, C. H. Q. Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sens.*, vol. 12, no. 21, 2020.

[36] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[37] Y. Zheng, M. Yang, M. Wang, X. Qian, R. Yang, X. Zhang, and W. Dong, "Semi-supervised adversarial semantic segmentation network using transformer and multiscale convolution for high-resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 8, 2022.

[38] S. Meng, Z. Di, S. Yang, and Y. Wang, "Large-scale weakly supervised learning for road extraction from satellite imagery," *ArXiv*, vol. abs/2309.07823, 2023.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.(CVPR)*, 2016, pp. 770–778.

[40] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops(CVPRW)*, 2018, pp. 172–17 209.

[41] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[42] Q. Zhu, Y. Zhang, L. Wang, Y. Zhong, Q. Guan, X. Lu, L. Zhang, and D. Li, "A global context-aware and batch-independent network for road extraction from vhr satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 353–365, 2021.

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

[44] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.(VCIP)*, 2017, pp. 1–4.

[45] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops(CVPRW)*, 2018, pp. 192–1924.

[46] 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.(ECCV)*, 2018.

[47] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, and S. Zhang, "Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist*, 2022, pp. 528–538.

[48] H. Lee and W. ki Jeong, "Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist*, 2020.

[49] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high- and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, 2021.

[50] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "Adaptmatch: Adaptive matching for semisupervised binary segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

Jingtao Hu is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing, computer vison and machine learning.

Qiang Li is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link https://crabwq.github.io/