Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss

Xianwei Zheng \textsuperscript{a,}\textsuperscript{*}, Linxi Huan \textsuperscript{a}, Gui-Song Xia \textsuperscript{a,b}, Jianya Gong \textsuperscript{a,c}

\textsuperscript{a} The State Key Lab. LIESMARS, Wuhan University, Wuhan, PR China
\textsuperscript{b} School of Computer Science, Wuhan University, Wuhan, PR China
\textsuperscript{c} School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, PR China

\textbf{ARTICLE INFO}

**Keywords:**
Semantic segmentation
Convolutional neural network (ConvNet)
Edge-aware loss

**ABSTRACT**

Parsing very high resolution (VHR) urban scene images into regions with semantic meaning, e.g. buildings and cars, is a fundamental task in urban scene understanding. However, due to the huge quantity of details contained in an image and the large variations of objects in scale and appearance, the existing semantic segmentation methods often break one object into pieces, or confuse adjacent objects and thus fail to depict these objects consistently. To address these issues uniformly, we propose a standalone end-to-end edge-aware neural network (EaNet) for urban scene semantic segmentation. For semantic consistency preservation inside objects, the EaNet model incorporates a large kernel pyramid pooling (LKPP) module to capture rich multi-scale context with strong continuous feature relations. To effectively separate confusing objects with sharp contours, a Dice-based edge-aware loss function (EA loss) is devised to guide the EaNet to refine both the pixel- and image-level edge information directly from semantic segmentation prediction. In the proposed EaNet model, the LKPP and the EA loss couple to enable comprehensive feature learning across an entire semantic object. Extensive experiments on three challenging datasets demonstrate that our method can be readily generalized to multi-scale ground/aerial urban scene images, achieving 81.7\% in mIoU on Cityscapes Test set and 90.8\% in the mean F1-score on the ISPRS Vaihingen 2D Test set. Code is available at: \url{https://github.com/geovsion/EaNet}.

1. Introduction

Semantic segmentation of urban scene images aims to locate objects at the pixel-level and assign them with categorical labels, which supports a wide range of urban applications, such as urban mapping and 3D modeling, autonomous driving, urban land cover classification and change detection (Zhu et al. 2017; Marcos et al. 2018; Zhao et al., 2018). However, as a dense pixel-wise classification task, semantic image segmentation faces big challenges in urban areas, due to the volume of detailed information contained in very high resolution (VHR) images and the large variations in the scale and appearance of objects. Large numbers of image details hamper the extraction of features relevant to global structure and semantic information of urban objects. Meanwhile, objects with large-scale variation frequently found in an image, such as large buildings and small cars, create difficulties when balancing the segmentation quality of images containing diverse kinds of objects varying in size. Moreover, the existence of many confusing categories, like trees and meadows, or similar objects with diverse appearances like cars, makes it hard to realize intra-class unification and inter-class discrimination simultaneously, when parsing urban scenes.

Extensive investigations have been presented for the challenging urban scene parsing task based on convolutional neural networks (ConvNets) (Yang et al. 2018; Yu et al., 2018; Zhao et al., 2018), due to the ability of ConvNets in hierarchical features learning and rich context capturing (Chen et al., 2018). In particular, ConvNets based on fully convolutional neural network (FCN) have become the mainstream approach for urban scene parsing with the success of the first end-to-end FCN for semantic segmentation (Long et al. 2015). However, the powerful ConvNets capability for abstraction in data-driven learning tasks creates two technical hurdles: imbalanced attention to multi-scale objects and loss of detail during encoding. Targeting these two issues, much effort has been devoted to the improvement of semantic segmentation (Liu et al., 2018; Yu et al., 2018).

In urban scene semantic segmentation, when there is scale variance in the objects found in an image, a neural network with an inappropriate receptive field size will give unbalanced attention to differently sized...
objects. A neural network with small view field will pay more attention to small things and divide the larger objects into fragments, while one with a large view field will ignore details and fail to separate small adjacent objects. Common solutions for multi-scale object segmentation focus on receptive field enlargement (Chen et al., 2018; Zhao et al., 2017). Many methods were developed with image pyramids (Zhao et al., 2018) or extra subnetworks (Yang et al., 2018), but such methods are time-consuming. A more popular way is to deploy a spatial pyramid pooling (SPP) module in the network architecture (Chen et al., 2018; Yuan and Wang, 2018; He et al. 2019). However, the current SPPs have difficulty in capturing relational information between long-range features while retaining continuous between neighboring features (Wang et al., 2018), due to inappropriate receptive field size design. Thus, when balancing segmentation quality of multi-scale urban objects, large objects still tend to be divided into fragments.

Another inevitable problem in urban scene semantic segmentation with ConvNets is detail degradation caused by downsampling. Detail degradation affects the accurate localization of objects at the pixel level, leading to blurry object boundaries. To tackle this problem, numerous methods have concentrated on enhancing the sensitivity of a model to boundary information. One way is to employ post-processing techniques such as a conditional random field (CRF) (Paisitkriangkrai et al. 2015; Sherrah 2016; Chen et al., 2018), which comes with high computational costs. The other relies on applying an extra edge extraction sub network (Cheng et al. 2017; Liu et al., 2018) or even an individual edge detection model like HED (Xie and Tu 2015; Marinis et al. 2018) to merge boundary information during segmentation. However, employing extra edge detectors will increase model complexity and require more training parameters. Moreover, the edge detectors used in these methods only learn edge features with pixel-level cross entropy loss (CE loss) and is independent to the semantic feature learning of an object, which lead to an incomplete learning across an entire object.

In this paper, we propose an edge-aware neural network (EaNet) for precise semantic segmentation of urban scenes. For the basic architecture of EaNet, we deploy a balanced encoder and decoder structure with skip pathways. To address the aforementioned two issues in a unified framework, we appended a couple of modules, i.e., large kernel pyramid pooling (LKPP) and Dice-based edge-aware loss function (EA loss) on the top of the encoder and the decoder of EaNet, respectively. The LKPP captures rich context information at multiple scales and builds strong continuous relations between long-range and neighboring features, by constructing several branches with different densely extending receptive field sizes. It effectively strengthens semantic unification inside objects to prevent them from being segmented into fragments. Moreover, the EA loss optimizes segmentation predictions via a standard cross entropy loss, and learns edge information directly from the segmentation prediction map using Dice-based edge loss. In this way, the EA loss module can work at both pixel- and image-level with no extra training parameters, which is more efficient and effective than many existing solutions for object boundary learning. By integrating the LKPP and the EA loss in a single one-stream EaNet model, the two modules can directly communicate through the forward and backward propagation, which enables a more comprehensive learning of semantic objects than many existing methods.

The EaNet is standalone and elegant, and has high generalization ability even with very large-scale urban scene data. Moreover, the two proposed modules, LKPP and EA loss, can be easily applied to other FCN frameworks. The main contributions of this work are follows:

- We propose a simple yet effective edge-aware neural network (EaNet) for a comprehensive learning of semantic objects.
- A LKPP module is proposed to densely capture multi-scale rich context with strongly continuous feature relations, and thus robustly segments multi-scale urban objects with high intra-class consistency.
- The EA loss module refines object boundaries directly from segmentation prediction at both pixel- and image-level, which significantly improves discrimination of confusing urban objects.
- The module provides a new loss function for simultaneous semantic category and edge structure learning, which is superior to existing combined solutions.
- We validate the proposed EaNet on three datasets with very different characteristics, i.e., Cityscapes, ISPRS Vaihingen 2D and WHU Aerial Buildings. We show that EaNet can be highly generalized to multi-scale ground/aerial urban scene data, achieving competitive performance.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. The architecture of EaNet and its components are detailed in Section 3. The performance of the two general modules and the complete EaNet is evaluated in Section 4. Some conclusions are drawn in Section 5.

2. Related work

Extensive works have been presented on urban scene semantic segmentation employing ConvNets, both in the field of computer vision and remote sensing (Zhu et al., 2017; Chen et al., 2018). In this section, we briefly review the works most relevant to the two technical hurdles in urban scene parsing, i.e., imbalanced attention to multi-scale objects and loss of boundary detail during encoding.

2.1. Multi-scale object segmentation

Object scale variance frequently occurs in urban scene images, and seriously affects the learning ability of deep networks in semantic segmentation. Many researchers have focused on enhancing the robustness to scale variance by view field enlargement and multi-level feature fusion. The approaches include image pyramid, extra sub-network, the encoder-decoder structure, and the SPP architecture techniques.

The image pyramid method parallelizes networks to learn features from inputs of several scales, where the long-range context and small object details become more prominent at different input sizes. Most image pyramid algorithms feed images at multiple scales to a network and merge the feature maps from all scales (Chen et al. 2016; Zhao et al., 2018). The extra sub-network approach performs multi-scale feature detection by cascading a feature extractor on top of the main network. Some methods add a DenseCRF (Krahenbühl and Koltun 2011) to model region similarities (Paisitkriangkrai et al. 2015; Jiang et al. 2017; Chen et al., 2018), and others have adopted sequential convolutional layers to extract long-range information (Liu et al. 2016; Yu and Koltun 2016). However, approaches built on image pyramids and extra sub-networks always involve high computational complexity and cannot effectively combine multi-level information.

An encoder-decoder framework equipped with spatial pyramid pooling (SPP) is the most common way for multi-scale object detection to efficiently extract hierarchical contexts and fuse the multi-level information. The encoder-decoder framework merges low-level and high-level features by skip pathways. Typical methods like U-Net (Ronneberger et al. 2015) perform feature aggregation via direct feature concatenation, while models like RefineNet (Lin et al. 2017), U-Net++ (Zhou et al. 2018) and ExFuse (Zhang et al. 2018) incorporate a complicated skip connection scheme between the encoder and the decoder. Networks can capture more context features by multiple effective receptive fields by embedding an SPP module into the encoder-decoder framework. Specifically, many variants of SPP have been designed for semantic segmentation achieving promising performance results. For example, the SPP module in PSPNet is composed of convolutions with kernels of four varying sizes (Zhao et al., 2017), while the atrous spatial pyramid pooling (ASPP) module proposed in DeepLab v2 groups parallel atrous convolution layers with different dilation rates (Chen et al., 2018). However, the current SPPs still face limitations. The SPP with standard convolution has difficulty in expanding the receptive
field by a large kernel size, otherwise will result in a large number of parameters. The SPP with small kernels (e.g., ASPP) lacks enough connection between neighbor features and the gridding problem, which happens when the view field is enlarged by a dilated convolutional layer (Wang et al. 2018).

In this work, we build EaNet on a balanced encoder and decoder structure with simple skip pathways. To overcome the drawbacks of the currently used SPPs, we constructed the LKPP module with several branches of different densely extending receptive fields, which can capture richer context information at multiple scales. Specifically, in the LKPP module, we propose hybrid asymmetric dilated convolution (HADC) blocks to continuously model relations among neighbor features with large kernels at a relatively low computation cost, which is beyond the capacity of the previous reported methods.

3. Architecture of the proposed EaNet

In this section, we discuss the architecture of the proposed EaNet and its two major components in detail, starting with an overview of the EaNet workflow in general.

3.1. Overview

The overall workflow of the proposed EaNet is shown in Fig. 1. EaNet takes an RGB image as input, and outputs a semantic segmentation prediction with an edge map. The basic architecture of EaNet is deployed in a balanced encoder and decoder structure, with a transfer layer embedded in each skip pathway. The two modules, LKPP and EA loss are appended on top of the encoder and the decoder, respectively. In the proposed EaNet architecture, the LKPP provides rich context information and fine-grained detail preservation during forward propagation, EaNet simultaneously semantic segmentation optimization and boundary refinement by the EA loss; while the EA loss offers geometry-aware guidance in the backpropagation for the LKPP to learn discriminative semantic and geometric features in a both pixel- and image-level manner. The LKPP and the EA loss couple to enable high-quality semantic segmentation with boundary refinement in a unified one-stream framework, i.e., the EaNet. In the following subsections, we will detail the LKPP and the EA loss module.

3.2. Rich context capture with large kernel pyramid pooling

Objects in urban scenes are usually characterized by a large variance of scale and such issue often makes it difficult to achieve unified semantic labeling at all scales. To fix this problem, a neural network must be adept at rich context information learning to satisfy the demand of multi-scale object recognition. In this section, we detail the rich context capture with the proposed large kernel pyramid pooling (LKPP).

3.2.1. Dilated convolution

For rich context extraction, a natural choice is to design neural network with receptive fields at diverse sizes. In current ConvNet models, SPP is a commonly used module to endow a neural network with various fields of view (Zhao et al. 2017; Chen et al., 2018), while many SPP variants apply sequential dilated convolutions. A 2-D dilated convolution with filter $w$ of size $K \times K$ is defined as:

$$Y[i,j] = \sum_{k_i} \sum_{k_j} X[i + r \cdot k_i,j + r \cdot k_j] \cdot w[k_i,k_j]$$

where $X$ represents a feature map and $Y$ is the corresponding output. $[i,j]$ and $[k_i,k_j]$ denote the location in the output $Y$ and filter $w$, respectively. The dilation rate $r$ corresponds to the stride. By changing the dilation rate, the receptive field size of its filters can be effectively enlarged.

Although dilated convolution was devised for receptive filed enlargement, the ‘holes’ in standard dilated convolution can disconnect and impair the relationship between local features resulting in gridding effect (Wang et al. 2018). As can be seen from Fig. 2(a), many pixels are ignored when using dilated convolution because the receptive field appears like a checkerboard, meaning much of the information from the input is discarded. An example of the gridding effect in urban scene parsing can be seen in Fig. 2(b): from left to right are input image, the ground truth and the predicted segmentation result, respectively. It is clear that the buildings in the predicted result were severely affected by the gridding effect, and segmented into irrelevant pieces.

In summary, most of existing methods perform edge refinement through a complex model with extra edge sub-networks and train their models with pixel-level loss function, which cannot directly optimize object boundaries in semantic segmentation prediction and lack object geometry guidance during training. Considering that segmentation labels and predictions contain rich and fine contour details, we developed a novel EA loss module in this work. The EA loss module is appended on top of the decoder, to fully exploit direct edge information source without introducing additional training parameters. The EA loss module deploys an image-level Dice loss for edge feature learning rather than pixel-level cross entropy loss, in which case EA loss can force the network to refine object boundaries with a truly built-in awareness of object geometry structure.
3.2.2. Large kernel pyramid pooling

To solve the aforementioned problem, we propose a new SPP module for rich multi-scale context extraction with high-quality detail preservation. As this new SPP module is equipped with large kernels based on a hybrid asymmetric dilated convolution (HADC), we name it as large kernel pyramid pooling (LKPP) to highlight the spatial pyramid pooling structure and the large kernel size. The main idea of LKPP is to capture richer context information with multiple branches, and each branch features a different receptive field size. As illustrated in Fig. 3, our LKPP module consists of two HADC blocks, a normal expanding block (Wang et al. 2018), a global context unit and a skip pathway (1 × 1 conv).

In LKPP, the most important part is the HADC block, which contains three stacked dual-direction layers that consist of two asymmetric dilated convolution (ADC) sub-layers. We define an ADC sub-layer as follows:

\[ Y[i, j] = \sum_{k_1}^{K_1} \sum_{k_2}^{K_2} X[i + r_1 \cdot k_1, j + r_2 \cdot k_2] \cdot w[k_1, k_2] \]  

where the size of filter \( w \) is \( K_1 \times K_2 \) and \( K_1 \neq K_2, r_1 \) and \( r_2 \) denote the dilation rates for the two dimensions of filter \( w \). We further pose a restriction to encode local relationship information from both two dimensions as \( \min(K_1, K_2) > 2 \). With this restriction, we regard the ADC layer as a substitute from standard dilated convolution. Unlike the \( 1 \times K \) (or \( K \times 1 \)) asymmetric convolution layer proposed by Szegedy et al. (2016), our ADC layer is able to capture richer region relationships because of the restriction above. Moreover, we use an ADC layer to replace a standard convolution layer, while Szegedy et al. (2016) stacked two asymmetric convolutions to factorize a standard convolution.

The ADC makes an effective trade-off between learning ability and
parameter amount. Given a filter with $K_1 = 3$ and $K_2 = 5$, the region covered by a $3 \times 5$ filter is 66.7% larger than a $3 \times 3$ filter, while the computational cost of a $3 \times 5$ filter is 40% cheaper than a $5 \times 5$ filter.

For more comprehensive scene context information, we combine two ADCs that have complementary kernels in terms of $x$ and $y$ directions (one with kernel size $K_1 \times K_2$ and the other with $K_2 \times K_1$) in a parallel way (we also tried the cascade configuration and found that the parallel version is more compatible with EaNet in our experiment), named the dual-direction layer, as illustrated in Fig. 4.

In practice, the dual-direction layer explores context with different attention to each direction, and then fuse the features from two directions by element-wise addition. In this case, the dual-direction layer learns a representation using a cross-shaped receptive field. As demonstrated by Ding et al. (2019), weights at the corner of a square filter usually contribute the least information in local feature extraction. Hence, the cross-like receptive field of our dual-direction layer reduces the influence of redundant information in representative feature learning. To enlarge receptive field for more context information, we constructed a HADC block with multiple dual-direction layers for dense receptive field enlargement. For different dual-direction layers, the dilation rates were designed to guarantee the relational continuity of neighbor features; thus, the gridding problem induced by ‘holes’ in standard dilated convolution are countered. Given an HADC block with three dual-direction layers $T_1, T_2, T_3$, and a feature map $X$, a HADC can be formulated as follows:

$$HADC(X) = T_3 \circ T_2 \circ T_1(X);$$

in which, $\forall n, m \in \{1, 2\}$ and $n \neq m, F(K_n, K_m, r_{i1}, r_{i2})$ denotes an ADC with kernel size of $K_n \times K_m$, $r_{i1}$ and $r_{i2}$ are the dilation rates for the two directions of a kernel, and $\circ$ represents the convolution operation. To prevent the ‘gridding’ effect that violates semantic consistency of neighboring features, the dilation rates in an HADC are set according to the following ‘completeness’ rules:

$$r_{i1}^1 = r_{i2}^1 = 1;$$

$$\max(|r_{i1}^n - 2r_{i1}^n|, r_{i2}^n) \leq K_n, \forall n \in \{1, 2\}$$

These rules guarantee that the receptive field of an HADC block will fully cover a large region without any missing pixels, and thus erase the gridding phenomenon.

In LKPP, we also insert a normal expanding module to compensate for information discarded by the HADC block. This normal expanding module is made by chaining three $3 \times 3$ standard symmetric convolutions, whose dilation rates also obey the ‘completeness’ rules. Meantime, a skip connection and a global context unit are applied in the LKPP.
skip connection is characterized by a $1 \times 1$ convolution to reuse the original signals and prevent gradient vanishing, and the global context unit is set to incorporate global contextual information. All these branches enable the LKPP module to perform simultaneous exploration of multi-scale regional features and global context information extraction effectively, without gridding.

3.3. Boundary refinement with a Dice-based edge-aware loss function

In urban scene semantic segmentation, over-smoothing of object boundaries is a common problem that affects segmentation quality. Over-smoothing can impair the boundary between two adjacent objects, making it difficult to distinguish confusing categories (e.g., the low vegetation and the tree). To alleviate boundary degradation, many methods rely on post-processing or independent HED-based edge detection (Cheng et al. 2017; Yu et al., 2018), which either have high computational cost or introduce numerous training parameters. More detection (Cheng et al. 2017; Yu et al., 2018), which either have high computational cost or introduce numerous training parameters. Moreover, the pixel-based cross entropy loss used in these methods pays little attention to the entire contour geometry of objects.

To address these problems, we propose a general Dice-based edge-aware loss (EA loss) module. EA loss module consists of a standard cross entropy loss function for semantic segmentation and a dice edge loss function for holistic contour fitting, as illustrated in Fig. 5. It localizes the object contours directly from the predicted segmentation maps using a simple edge extractor, which needs no extra sub network for edge training and prediction. In the following, we detail the two main parts of EA loss module, semantic edge map prediction and edge-aware loss function.

3.3.1. Semantic edge map prediction

Recall the cross entropy loss (CE loss) as defined follows:

$$L_{ce}(\hat{Y}, Y|I, \Theta) = -\sum_{i,j} p(y_{ij} = k|I) \log \hat{y}_{ij} = k|I, \Theta)$$

where $I, \Theta, \hat{Y}$ denote the input image, the model parameters, and the predicted label, respectively. $p(y_{ij} = k|I)$ and $p(y_{ij} = k|I)$ refer to the probability of a pixel at the location $(i,j)$ of the prediction $\hat{Y}$ and the label $Y$ that belongs to the object denoted by class $k$. The CE loss function does not directly consider the relationship between different objects. It treats every pixel separately in the semantic map without linking pixels to their neighbors, and pays equal attention to edge pixels and non-edge pixels. Such two attributes make the cross entropy function weak in structure learning, usually leading to blurry object boundaries. To strengthen the structure capturing ability, an edge loss function that works at image-level, is necessary in addition to the CE loss function.

The key insight here is that the segmentation map contains rich edge information that can be utilized for direct boundary refinement in segmentation prediction. To accomplish this goal, we appended a simple edge extractor on top of the segmentation network by constructing a squash model upon the Laplacian operator. Defined as the divergence of the gradient of the probability of a pixel at the location $(i,j)$ of the prediction $\hat{Y}$ and the label $Y$ that belongs to the object denoted by class $k$.

The key insight here is that the segmentation map contains rich edge information that can be utilized for direct boundary refinement in segmentation prediction. To accomplish this goal, we appended a simple edge extractor on top of the segmentation network by constructing a squash model upon the Laplacian operator. Defined as the divergence of the gradient of the probability of a pixel at the location $(i,j)$ of the prediction $\hat{Y}$ and the label $Y$ that belongs to the object denoted by class $k$.

$$\Delta f = \frac{\partial f}{\partial x^2} + \frac{\partial f}{\partial y^2}$$

where $f$ refers to an image, and $x, y$ are the two coordinate directions of $f$. The output of the Laplacian operator performed on the predicted semantic segmentation maps are termed gradient information maps $G = \{g_1, ..., g_C\} \in \mathbb{R}^{H \times W \times C}$, in which $g_1, ..., g_C$ are the gradient information maps for every object class. Since the Laplacian operator generates a gradient information map $g_k \in G$ for every category, it is natural to directly derive semantic edge maps $E = \{e_1, ..., e_C\} \in \mathbb{R}^{H \times W \times C}$ via an activation function from $G = \{g_1, ..., g_C\}$, and refine $E = \{e_1, ..., e_C\}$ with semantic edge labels (derived from the semantic ground truth via the Laplacian operator) (Liu et al., 2018). However, such rough refinement will result in semantic conflict and requires more GPU memory. A border between two objects can be classified as either class, but the rough refinement strategy forces the network to assign a specific class to the border, which actually wastes part of the learning ability on an unnecessary fitting task. Additionally, every semantic edge map $e_k \in E$ needs a corresponding semantic edge label for the original image size for loss computation as it needs $C$ times more GPU memory to restore training labels than a single edge ground truth.

To achieve a class-agnostic boundary refinement with only one edge label map, we use a function to convert the gradient information maps into a single edge probability map. Given the gradient information maps $G = \{g_1, ..., g_C\} \in \mathbb{R}^{H \times W \times C}$, the gradient information vector $\bar{g} \in \mathbb{R}^{1 \times 1 \times C}$ at every position $(i,j)$ records the gradient information for

![Fig. 5. The components of the edge-aware module.](image-url)
every class. To obtain a single edge probability map, we apply a squash function (7) to further calculate the gradient intensity, and compress the intensity value into $[0, 1]$. 

$$\hat{e}_q = \operatorname{squash}\left(\frac{\|g_q\|^2}{\|g_q\|^2 + \alpha}\right)$$  (7)

In (7), $g_q$ and $\hat{e}_q$ denote the gradient information vector and edge probability at $(i,j)$, respectively. $\|\cdot\|$ is the $L_2$-norm, and $\alpha$ is a hyper-parameter that controls the model’s sensitivity to object contours. If $\|g_q\| > \alpha$, the edge probability $\hat{e}_q$ at location $(i,j)$ will be greater than 0.5, meaning that the pixel at $(i,j)$ is more likely an edge point. Hence, when $\alpha$ is small, pixels even with a weak response (i.e. small $\|g_q\|$) will be regarded as part of the edges, and a slight oscillation in the predicted segmentation map may result in unexpected edges predicted inside an object. An interesting thing is that, the unexpected edges are suppressed during training (as shown in Fig. 6); hence, local differences between the gradients of neighboring pixels around unexpected edges are smoothed. In our experiments, we found that setting $\alpha$ to 1 achieves an optimum balance between intra-class unification and inter-class discrimination.

By applying the squash function on the gradient information maps $G$, the proposed edge extractor obtains a class-agnostic edge prediction from a semantic segmentation map. In this case, the edge extractor can be applied in any semantic segmentation network $F$ to predict an edge probability map $\hat{E} = [e_q]_{H \times W}$ prediction. Given an input image $I$, the whole procedure represents a complete function as: $\hat{E} = G_F(I)$.

### 3.3.2. Dice-based edge-aware loss function

With the predicted edge probability map $\hat{E}$ and the edge label $E = [e_q]_{H \times W}$, boundary refinement is expressed as a Dice coefficient maximization problem defined by (8):

$$\hat{\theta} = \arg \max_{\theta} \text{Dice}(\hat{E}, E | I, \theta) = \arg \max_{\theta} \frac{2 \sum_{i=1}^{H} \sum_{j=1}^{W} e_q \hat{e}_q}{\sum_{i=1}^{H} \sum_{j=1}^{W} e_q^2 + \sum_{i=1}^{H} \sum_{j=1}^{W} \hat{e}_q^2}$$  (8)

where $\theta$ is the parameters of the segmentation network $F$. To implement SGD during training, the final edge loss is constructed as:

$$L_{\text{edge}}(\hat{E}, E | I, \theta) = 1 - \text{Dice}(\hat{E}, E | I, \theta)$$  (9)

where the minimization of (9) equals to the maximization of $\text{Dice}(\hat{E}, E)$.

Different from other commonly used pixel-level edge loss functions (e.g., cross entropy, L1 and L2 loss) that treat the predicted pixels as independent individuals, the Dice loss deployed in the EA loss function builds an image-level association across all points. In fact, since the Dice coefficient measures the similarity of two sets, the maximization of $\text{Dice}(\hat{E}, E)$ can be thought as an image-level procedure. It holistically fits the geometric structure of the predicted edge set $\hat{E}$ to that of the true edge set $E$ as closely as possible; the parameters $\theta$ are thus updated with more attention to the geometric details of the objects in an image. Specifically, the Dice coefficient edge loss function avoids sample imbalance between edge points and non-edge points. The final EA loss function in EaNet is a combination of the standard cross-entropy and edge loss functions, and defined as follows:

$$L_{\text{EA}}(I, Y, E) = L_{\text{ce}}(\hat{Y}, Y | I, \theta) + \lambda_1 L_{\text{edge}}(\hat{E}, E | I, \theta) + \lambda_2 \|	heta\|_2$$  (10)

in which, $I, Y, E$ refer to the input image, the semantic segmentation ground truth, and the edge label, respectively. $\lambda_1$ denotes a penalty parameter for balancing the standard cross-entropy term and edge fidelity term, $\lambda_2$ is the decay weight, and $\|	heta\|_2$ is the $L_2$-norm of all the model parameters. In our experiment, we set $\lambda_1$ as 0.5 to prevent the model from learning distinctive features that might weaken the ability of the algorithm to generalize; and we added the regularization term $\|	heta\|_2$, to prevent over-fitting.

As a prerequisite for holistic end-to-end neural network training, the

![Groundtruth](image1.png)

![Before smoothing](image2.png)

![After smoothing](image3.png)

Fig. 6. An example of suppressing the unexpected edges to smooth gradients of neighboring pixels around them.
EA loss function is differentiable. Its gradient computed by semantic segmentation probability prediction at position \( (i,j) \) for the \( k \)th class \( p^k_{ij} \), is as follows:

\[
\begin{align*}
\frac{\partial L_{edge}}{\partial p^1_{ij}} &= \frac{\partial \text{Dice}}{\partial p^1_{ij}} = 2 \sum_{r=1}^{1} \sum_{s=1}^{1} \frac{\partial e_{ij+r,s}}{\partial p^1_{ij}} (1 - \overline{e}_{ij})^2 / \alpha, \text{ if } r = 0 \text{ and } s = 0; \\
&= 2 \overline{e}_{ij+r,s} (1 - \overline{e}_{ij+r,s})^2 / \alpha, \text{ if } r \text{ or } s \in \{-1, 1\}; \\
&= 0, \text{ otherwise.}
\end{align*}
\]

From the gradient formulas for cross entropy and Dice loss, we can see that \( L_{edge} \) refines the semantic segmentation prediction by comparing individual pixels in the inference map and the ground truth. The term \( L_{edge} \) improves the segmentation quality by taking the local neighbors of a pixel, and the global context into consideration.

4. Experiments

We conducted experiments on three datasets, including a large-scale ground dataset, i.e., Cityscapes (Cordts et al. 2016), and two relatively small-scale aerial datasets, i.e., ISPRS Vaihingen 2D (Gerke 2014) and the WHU Aerial Building Dataset (Ji et al. 2018), in order to comprehensively test the learning capacity and generalizability of the proposed EaNet model. Ablation studies were conducted for the two general modules, i.e. LKPP and EA loss individually, to verify their efficacy when parsing urban scenes. In the following subsections, we will detail the dataset description, experimental settings and our experimental results on the three datasets.

4.1. Dataset and implementation details

4.1.1. Dataset

Cityscapes Dataset: The Cityscapes dataset is a large-scale ground dataset for urban scene parsing (Cordts et al. 2016), which contains 5000 ground natural RGB images collected from 50 different cities. Each image is \( 2048 \times 1024 \) pixels with 19 high quality annotated semantic classes. Samples in this dataset are split into three parts: a training set of 2975 images, a validation (val) set of 500 images, and a test set of 1,525 images.

ISPRS Vaihingen Challenge Dataset: This is an ISPRS 2D semantic labeling challenge benchmark dataset. The dataset includes 33 very high-resolution true orthophoto (TOI) images (GSD ~ 9 cm) at a \( 2500 \times 2000 \) pixel size, and two sets of auxiliary data, the Digital Surface Model (DSM) and Normalized Digital Surface Model data (NDSM). This dataset was officially split into 16 areas for training and 17 areas for testing. For the validation dataset, the setup is the same as (Volpi and Tuia 2016; Marcos et al. 2018; Ghassemi et al. 2019; Mou et al. 2019), in which 11 images are selected for training, and the remaining five images (image IDs: 11, 15, 28, 30, 34) were used for validation. We conducted the experiments on both the validation and test dataset, only TOP images were used.

WHU Aerial Building Dataset: The WHU aerial building dataset is a new open-source challenge benchmark for building detection (Ji et al. 2018). The WHU aerial dataset contains 8189 tiles seamlessly cropped into \( 512 \times 512 \) from aerial images at a 0.3-m spatial resolution, where over 187,000 building instances were labeled. These building samples vary in architectural types, color, size, and usage. All tiles in this dataset were officially split into three parts: a training set composed of 4736 images, a validation set consisting of 1036 images, and a test set including 2416 images.

4.1.2. Experimental settings

Our method was implemented using the Pytorch framework. Following Chen et al. (2018), the base learning rate was set to 0.01. A poly learning rate policy was employed, in which the initial learning rate was multiplied by \( (1 - \frac{iter}{total\ iter})^{0.9} \) during each iteration. All models in the experiments were trained with the SGD optimizer on NVIDIA GTX 2080ti GPUs. The momentum value was 0.9 and the weight decay value was 5e-4. Data augmentation techniques, including random scaling, random cropping, and random flip, were applied. During the training process, images and labels were randomly resized with scale factors from 0.5 to 2.0, and cropped into \( 789 \times 789 \) for Cityscapes, and 512 \( \times 512 \) for ISPRS Vaihingen 2D and WHU Aerial Building datasets. The training time was 60 k iterations for Cityscapes, and 40 k iterations for the two remaining datasets. The batch size was chosen according to the dataset, as detailed in the following subsections. We also applied multi-scale inference by averaging the segmentation probability maps computed at multiple scales, following Chen et al. (2018).

<table>
<thead>
<tr>
<th>Method</th>
<th>Inception</th>
<th>ASPP</th>
<th>LKPP</th>
<th>CE loss</th>
<th>EA loss</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>71.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>✓</td>
<td></td>
<td></td>
<td>73.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deeplab v3+</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>75.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deeplab v3+</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>75.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>76.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>71.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>78.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.3. Evaluation metrics

The performance of our EaNet model on different datasets was estimated by mean intersection over union (mIOU), precision, recall, F1-score, and overall accuracy (OA). The evaluation was based on an accumulated confusion matrix, from which mIOU, precision, recall, F1-score, and overall accuracy can be derived:

\[
\text{mIOU} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}; \\
\text{Precision} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k}; \\
\text{Recall} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + TN_k + FN_k}; \\
\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}; \\
\text{OverallAccuracy} = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k},
\]

where \( TP_k, FP_k, TN_k, FN_k \) denote the true positive, false positive, true negative and false negative pixels respectively, for object indexed as class \( k \).

4.2. Results on cityscapes dataset

The ablation experiment on Cityscapes dataset includes two parts, an evaluation of the LKPP and the EA loss modules, and a comparison with other state-of-the-art methods.

4.2.1. Ablation study for LKPP and EA loss

To test the LKPP and the EA loss modules independently, we inserted them into a network architecture the same as EaNet, as illustrated in Fig. 1. The ResNet50 network (He et al. 2016) was chosen as the backbone and the training batch was set to eight. We conducted experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>LKPP</th>
<th>EA loss</th>
<th>mIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>✓</td>
<td></td>
<td>75.27</td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>79.91</td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>76.84</td>
</tr>
<tr>
<td>EaNet</td>
<td>✓</td>
<td>✓</td>
<td>80.90</td>
</tr>
</tbody>
</table>

Table 2

Ablation study on Cityscapes val set with ResNet101 for the two modules, LKPP and EA loss. (unit:%).

Fig. 7. Visualization results of the ablation study of LKPP module on Cityscapes val set.

Fig. 8. Visualization results from the ablation study of EA loss on the Cityscapes val set.
for further validation. As shown in Table 1, the LKPP module alone yields a mIoU score of 76.93%, outperforming the baseline by 5.78%.

comparison. Furthermore, the EA loss was not only applied to the comparison, as like LKPP, they are also two modules used for multi-scale context learning. Since the inception block has many variants, here the context learning. Since the inception block has many variants, here the
under different settings to verify the performance of LKPP and EA loss.

Table 3
Quantitative comparison with the state-of-the-art methods on Cityscapes dataset. † indicates training with train-fine for test set. ‡ indicates training with train-fine and val-fine for test set. Values in bold and underlined indicate the first and second place. (unit:%).

Table 4
Quantitative comparison results on ISPRS Vaihingen validation set, where the values in bold are the best. The short names of different methods are cited from

Table 5
Quantitative comparison results on ISPRS Vaihingen challenge 2D test set, where the values in bold are the best. The short names of different methods are cited from the challenge evaluation website. (unit:%).

The LKPP module also outperforms the inception module and the ASPP respectively by 3.84% and 1.55% in mIoU, revealing that the LKPP has a stronger ability for rich context learning. The EA loss alone yields a mIoU score of 71.56%, outperforming the baseline model by 0.41%. It is not surprising that the EA loss alone does not show a distinct rise in mIoU as the LKPP alone. The two probable reasons are that the few boundary pixels cannot affect much of the total mIoU, and the baseline model without a multi-scale context capture module does not provide sufficient information for refinement. The latter can be inferred by applying the LKPP module and the EA loss together, which gains over the LKPP by 1.1% in mIoU score. This means that the EA loss can gain more accuracy improvement with the rich context extracted by the LKPP module. The EA loss alone also boosted the performance of Deepab v3 + by 0.6%, which further demonstrates the superiority of the EA loss over the commonly used CE loss.

For a more comprehensive evaluation with a stronger backbone network, we also did an ablation study with ResNet101, and the results are listed in Table 2. From Table 2, it can be seen that the LKPP and the EA loss both can improve the performance of the baseline network, and the highest accuracy was achieved when the LKPP and the EA loss work together.

For an all-round comparison, the effects of LKPP and EA loss module are visualized in Fig. 7 and Fig. 8, which permit an intuitive, perceptual understanding of the advantages of the two modules. In the Cityscapes dataset, vehicles including cars, trains and so on, occupy a prominent position. They are divided into several types, which are very difficult to distinguish due to their similar appearance. In this situation, our LKPP module delivers coherent semantic annotation within an object, which is often violated when gridding occurs, as our EA loss function separates adjacent semantically distinct entities using fine-gained contour learning.

Fig. 7 illustrates that our LKPP module generates intact segmentation results for all vehicles with higher semantic unification than the baseline.
network and Deeplab v3+. As shown in the first row of Fig. 7, the baseline network correctly predicted a visually large bus close to the camera, but misclassified the two visually small buses farther away, as shown in the zoomed window. This happened because there is no multi-scale receptive field; therefore, the baseline network directed biased attention to multi-scale objects. By introducing ASPP, Deeplab v3+ detected buses near the camera and those at some distance away, revealing that ASPP has the ability to recognize multi-scale objects. Unfortunately, Deeplab v3+ was affected by the gridding problem and yielded partial misrecognition results and incomplete shape for the bus closer to the camera due to the sparse receptive field in ASPP. From the last column of Fig. 7, it can be seen that EaNet successfully segmented all the buses and cars with high intra-class semantic consistency. Our EaNet model overcomes the limitations of the baseline network and Deeplab v3+ with its dense hybrid receptive field. More concretely, our LKPP module is capable of extracting much richer multi-scale context features with continuous regional relations, thus avoiding segmenting one object into incomplete fragments. Comparative visual results of the EA loss are showcased in Fig. 8.

With regard to the performance of EA loss function, its robustness for inter-class discrimination is revealed by the qualitative results seen in Fig. 8, as the structural details of object are clearer and the boundaries between two similar objects are more distinct with EA loss, such as the truck and the car in the first row, and the person and the car in the second row. Particularly, in the third row of Fig. 8, the EaNet model with EA loss function detected the wall without wrongly classifying pixels as vegetation. In contrast, the networks without the EA loss function were affected by the shadow cast on the wall. Unlike the commonly used CE loss function, the EA loss function guides the network to learn features at both pixel- and image-level, which enables the network recognize objects using their geometrical structure. This permits the model to discern two neighboring objects with different semantics.

### 4.2.2. Comparing EaNet to state-of-the-art models

For comparisons with other state-of-art methods, the backbone network was built on ResNet101, with a batch size set to twelve on Cityscapes validation and test dataset, respectively. The results of our EaNet and other state-of-the-art methods are provided in Table 3 as a reference, and we only list the results trained with merely the fine labeled dataset provided by Cityscapes for comparability. The numeric results in Table 3 show that our method can perform favorably against many existing approaches. However, we emphasize that it is not appropriate to simply compare the results in Table 3, since these methods were trained with different settings, training sets and

![Fig. 9. Qualitative comparison results on ISPRS Vaihingen 2D challenge test set. The label includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow) and clutter/background (red). Enlarge the PDF version to >=200% to get a better visual effect for object boundaries.](image)

### Table 6

Quantitative comparison results with the state-of-art models on WHU Aerial Building dataset (unit: %).

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web-Net</td>
<td>ResNet101</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>88.76</td>
</tr>
<tr>
<td>SRI-Net</td>
<td>ResNet101</td>
<td>93.28</td>
<td>95.21</td>
<td>94.23</td>
<td>89.09</td>
</tr>
<tr>
<td>FastFCN</td>
<td>FCN</td>
<td>81.37</td>
<td>87.98</td>
<td>84.55</td>
<td>73.23</td>
</tr>
<tr>
<td>EU-Net</td>
<td>U-Net</td>
<td>95.10</td>
<td>94.98</td>
<td>95.04</td>
<td>90.56</td>
</tr>
<tr>
<td>SRI-Net-UC</td>
<td>ResNet101</td>
<td>93.69</td>
<td>95.07</td>
<td>94.51</td>
<td>89.23</td>
</tr>
<tr>
<td>DE-Net</td>
<td>ResNet</td>
<td>94.60</td>
<td>95.00</td>
<td>94.08</td>
<td>90.12</td>
</tr>
<tr>
<td>SIU-Net</td>
<td>U-Net</td>
<td>93.90</td>
<td>93.80</td>
<td>93.85</td>
<td>88.4</td>
</tr>
<tr>
<td>U-Net</td>
<td>U-Net</td>
<td>94.50</td>
<td>90.14</td>
<td>92.27</td>
<td>86.8</td>
</tr>
<tr>
<td>EaNet</td>
<td>ResNet101</td>
<td>96.09</td>
<td>94.63</td>
<td>95.35</td>
<td>91.11</td>
</tr>
</tbody>
</table>
backbones. For example, Deeplab v3+, DenseASPP, DUC and HRNetV2 are built on more powerful pretrained models; ICNet adopts a simple backbone and hence sacrifices the segmentation accuracy for inference speed; OCR uses auxiliary soft objet region detection loss. Therefore, the results just validate that the proposed EaNet can achieve competitive performance with other state-of-art methods with a simple general network architecture.

4.3. Results on the ISPRS Vaihingen dataset

We carried out experiments on the ISPRS Vaihingen dataset, to further evaluate the effectiveness and the generalization ability of our EaNet on areal urban scene images. As stated by Ghassemi et al. (2019), the ISPRS Vaihingen dataset is characterized by a small covariate shift between training and test sets, high performance can be easily made with specifically designed models, especially for those fuse orthophoto (TOP) images with auxiliary DSMs (Audebert et al. 2016; Sherrah 2016). In this part, we will show that our EaNet model using only TOP images as input and relying on a simpler architecture, can also outperform those specially designed models. The backbone was also ResNet101, the batchsize for this dataset was set to ten during the training stage.

The performance of EaNet was first evaluated on the validation dataset and compared with methods that adopted the same splitting schedule (see the Dataset part). The results are listed in Table 4. The numeric scores for the validation dataset show that our EaNet delivers robust performance, and exceeded other methods in the F1-scores and the overall accuracy (OA) by a considerable margin.

For a comprehensive evaluation, EaNet was also tested on the Vaihingen test set and compared with other methods. The quantitative results and the model details including the combination strategies of different methods are shown Table 5. As can be seen in Table 5, the proposed EaNet achieved the top performance on the category of imp suf, building and car, and showed a best Mean F1-score, even though EaNet model uses a simpler architecture than those specially designed models with complex structures. The numeric results in Table 5 show almost all the methods achieve promising performance on building categories; however, most existing methods perform poorly in car prediction. Especially, the performance of HUSTW5, showed a drastic difference between large and small objects (e.g., 96.1% on building but only 74.6% on car), revealing that HUSTW5 was weak in balancing the segmentation quality of multi-scale objects. In contrast, EaNet exceeded the others by nearly 32.7% to 5.9% on car category. This indicates that EaNet is more capable of balancing the segmentation quality of objects at diverse scales.

In Fig. 9, the effect of segmentation results on Vaihingen test set for the different methods are visualized for a better inspection. It can be seen that despite the use of DSM data, most of existing methods still fail to annotate the very large buildings with confusing texture coherently, as shown in the first, third and fourth rows of Fig. 9. Benefiting from the use of the LKPP and EA loss module, our EaNet model can obtain a coherent and accurately labeled result in these uneven regions that are hard to distinguish. In our EaNet, the LKPP and EA loss together learn the features for an entire semantic object, which are thus more capable of segmenting objects with complete structures. A more representative example can also be found in the second row of Fig. 9. Our EaNet successfully separated adjacent buildings at the top right region, where the shadow between the two buildings severely interferes with the predictions from other approaches. EaNet also draws out the complete shape of the building at the left bottom part with a clear boundary, where other methods yielded incomplete building shapes due to interruptions caused by different textures. All these results demonstrate that our EaNet is more robust to adjacent object confusion and effectively captures fine-structured objects in a most holistic fashion. In addition, the object boundaries generated by our EaNet are remarkably sharper than those from other methods, especially when it comes to regular objects, such as buildings. This phenomenon validates the effect of EA loss in edge refinement and geometrical structure constraint, which makes our EaNet competitive in accurate contour localization.

4.4. Results on WHU Aerial building dataset

Different from the previous two datasets, the WHU Aerial Building dataset is principally used for building detection, in which the ground truth includes only two classes, background and buildings. Since
buildings feature straight lines, the balance weight for the edge fidelity term in EA loss was increased from 0.5 to 1.0, for stricter geometric structure control. Furthermore, the training batch size for this dataset was twelve. The quantitative performance comparisons on the benchmark test on the WHU Aerial Building dataset are summarized in Table 6, including the results from Web-Net, SRI-Net, FastFCN and EU-Net as reported by Kang et al. (2019); SRI-Net-UC and DE-Net are reported by Liu et al. (2019); SIU-Net and U-Net are reported by Ji et al. (2018).

From Table 6, it is clear that EaNet achieves the highest prediction scores in terms of the recall, F1-score and IoU metrics. The high recall score indicates that EaNet successfully detects more buildings than other methods. There is large-scale variance in WHU Aerial Building dataset, so the high recall score also reveals the effectiveness of the proposed EaNet in handling multi-scale object detection. The F1-score and the IoU metric can take both the detection accuracy and completeness into consideration, and thus are more suitable for comprehensive performance evaluation than the recall and precision metrics. Hence, the high scores in F1-score and IoU tell that our EaNet model obtains both high detection accuracy and completeness.

As for the visualization results given in Fig. 10, it can be seen that there is not only a perceptual variance in building scales and appearance, but also a severe class imbalance between the buildings and background. Actually, buildings only occupy about 18.7% of the area covered by the training data. Even though interrupted by the background, the EaNet still achieved high performance in building extraction, and managed to tackle multi-scale issues effectively. As can be seen from Fig. 10, EaNet accurately localized both large and tiny buildings, and distinguished between two very near buildings instead of wrongly joining them together. Considering class imbalance, as shown in the first column, the EaNet also discerned all the buildings from background interference. Moreover, our EaNet model still profiled buildings with strongly defined boundaries, prominently revealed by the clear building contours in the second column of Fig. 10. All these results again demonstrate the two modules, LKPP and EA loss; have a strong ability to learn discriminative features that contribute to precise urban scene segmentation.

5. Conclusion

In this paper, we propose an edge-aware neural network (EaNet) with large kernel pyramid pooling for robust semantic segmentation in urban areas. Extensive ablation experiments show that the proposed EaNet can adapt to both ground and aerial urban scene images, and achieved excellent performance consistently on three benchmark datasets, i.e., Cityscapes, ISPRS Vaihingen, and the WHU Aerial Building datasets. Qualitative and quantitative analysis results verify that the two introduced modules, i.e., the LKPP and the EA loss modules, effectively and comprehensively learns semantic objects and thus enables precise urban scene semantic segmentation. Specifically, the individual ablation study for the two modules convincingly demonstrates that the LKPP module is capable of multi-scale object detection with high intra-class semantic consistency, and the EA loss can distinguish between confusing objects with clearly defined boundaries.

It is easy to design a specialized model with complex structures to gain high performance when using a small dataset with little domain shift, but difficult to maintain the same high-quality inference in terms of other more complex datasets. In our experiment, the EaNet was tested on three different datasets with extreme variation in data scales, image characteristics and kinds of classes. The performance evaluation on these datasets confirms the effectiveness and generality of our EaNet model equipped with the powerful LKPP and EA loss modules. More importantly, the proposed LKPP and EA loss modules can be applied in any semantic segmentation neural network for coherent labeling and accurate object structures. However, the EaNet mainly focuses on outdoor RGB scene images, which is inadequate to handle indoor scenes that contain severe occlusion and quantities of irregular instances and intricate structures. In the future work, we will try to combine EaNet with task-specific fusion modules to bridge the RGB and depth information for improved indoor scene parsing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Key Research and Development Program of China under Grant 2018YF0505401, the National Natural Science Foundation of China Project under Grant 41701445, 41873161, 42071370 and the Fundamental Research Funds for the Central Universities.

References


