International Journal of Applied Earth Observation and Geoinformation 136 (2025) 104309

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Small but mighty: Enhancing 3D point clouds semantic segmentation with U-Next framework

Ziyin Zeng [a,b], Qingyong Hu [c], Zhong Xie [a], Bijun Li [b], Jian Zhou [b], Yongyang Xu [a,*]

[a] School of Computer Science, China University of Geosciences, Wuhan, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[c] Academy of Military Sciences, Beijing, China

ARTICLE INFO

Dataset link: https://github.com/zeng-ziyin/U-Next

Keywords:
Point cloud
Semantic segmentation
Network architecture
Deep supervision

ABSTRACT

We investigate the problem of 3D point clouds semantic segmentation. Recently, a large amount of research work has focused on local feature aggregation. However, the foundational framework of semantic segmentation of 3D point clouds has been neglected, where the majority of current methods default to the U-Net framework. In this study, we present U-Next, a small but mighty framework designed specifically for point cloud semantic segmentation. The key innovation of this framework is to capture multi-scale hierarchical features. Specifically, we construct the U-Next by stacking multiple U-Net L^1 sub-networks in a dense arrangement to diminish the semantic gap. Concurrently, it integrates feature maps across various scales to proficiently restore intricate fine-grained details. Additionally, a multi-level deep supervision mechanism is introduced for smoothing gradient propagation and facilitating network optimization. We conduct extensive experiments on benchmarks, including the indoor S3DIS dataset, the LiDAR-based outdoor Toronto3D dataset, and the urban-scale photogrammetry-based SensatUrban dataset, demonstrate the superiority of U-Next. The U-Next framework consistently exhibits significant performance enhancements across various benchmarks and baselines, demonstrating its considerable potential as a versatile point-based framework for future endeavors. The code has been released at https://github.com/zeng-ziyin/U-Next.

1. Introduction

With the considerable advancements in 3D sensors, the analysis of 3D point clouds has garnered increasing attention recently (Guo et al., 2020) and has been applied in indoor navigation, autonomous driving, and smart cities (Behley et al., 2019; Rusu and Cousins, 2011; Blanc et al., 2020; Zeng et al., 2024a). Serving as a pivotal facet of 3D scenes understanding, semantic segmentation of 3D point cloud is aimed at assigning a distinct semantic label to each point based on its geometrical structure, and has garnered significant focal point of extensive researches. However, attaining fine-grained semantic scene parsing poses a persistent and formidable challenge since the acquired point clouds are naturally orderless, irregular, and unstructured, coupled with large-scale points, varying-size objects, and non-uniform distributions.

Recently, the pioneering research work PointNet (Qi et al., 2017) has been proposed to directly learn from unstructured 3D point clouds, while with high efficiency and encouraging performance on several downstream tasks. Later, deep learning on point cloud analysis (Guo et al., 2020) has been extensively explored. A plethora of sophisticated neural architectures (Qi et al., 2018; Yan et al., 2020; Hu et al., 2021b; Ao et al., 2022; Qiu et al., 2021; Hu et al., 2022a; Wang et al., 2019a; Nie et al., 2022; Thomas et al., 2019; Graham et al., 2018a; Zeng et al., 2024a) have been proposed, leading to consistent improvements in performance on large-scale 3D point cloud benchmarks (Armeni et al., 2016; Hackel et al., 2017; Behley et al., 2019; Hu et al., 2022b; Chen et al., 2022). Despite the promising progress achieved, existing approaches have primarily focused on representation learning, *i.e.,* learning an effective representation for downstream tasks of 3D point clouds, including point-based (Qi et al., 2017, 2018; Hu et al., 2020; Qian et al., 2022; Zhao et al., 2019), projection-based (Boulch et al., 2017; Kundu et al., 2020), voxel-based (Tchapmi et al., 2017; Graham et al., 2018a; Choy et al., 2019), and point-voxel-based techniques (Tang et al., 2020; Liu et al., 2019). Furthermore, several methods have extensively explored various local feature aggregation mechanisms (Hu et al., 2020; Yan et al., 2020; Li et al., 2019) and efficient neural network architectures (Graham et al., 2018a; Choy et al., 2019). However, few endeavors were made to improve the segmentation framework in the field of 3D point cloud learning. Most

* Corresponding author.
E-mail address: yongyangxu@cug.edu.cn (Y. Xu).

https://doi.org/10.1016/j.jag.2024.104309
Received 22 December 2023; Received in revised form 21 October 2024; Accepted 2 December 2024
Available online 13 December 2024

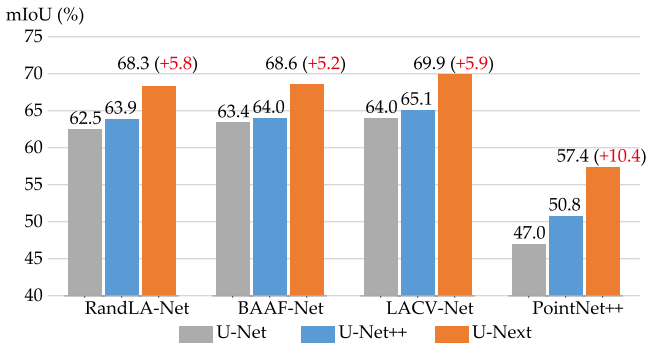


Fig. 1. Performance of various algorithms on S3DIS dataset (Area 5) with different frameworks. Our U-Next framework demonstrated a notable advantage to both U-Net and U-Net++.

existing approaches tend to indiscriminately employ the widely-used U-Net architecture, while overlooking its inherent limitations, which have been extensively documented in the context of 2D natural and medical images.

As one of the most widely used neural architectures, U-Net (Ronneberger et al., 2015) is built upon the encoder–decoder structure, and the key is to use skip connections to combine intermediate latent feature maps from encoder and decoder. Although this architecture appears symmetrical and natural, the combined same-scale feature maps are actually far from semantically similar (*shallow* and *low-level* features from encoder + *deep* and *semantic* features from decoder), and there is no solid theory to prove such a combination is optimal (Zhou et al., 2019). On the other hand, the optimal depth of the network is also apriori unknown, which requires manual architecture tuning. To this end, a handful of approaches (Xiao et al., 2018; Guan et al., 2019; Chen et al., 2021; Wang et al., 2022a; Xiao et al., 2018; Guan et al., 2019) have started to explicitly improve the U-Net framework, including U-Net++ (Zhou et al., 2019) with nested, dense re-designed skip pathways, UNet3+ (Huang et al., 2020) with full-scale skip connections, and TransUNet (Chen et al., 2021) with transformer as encoder, *etc*. Although proven to be highly effective in the segmentation of medical images, it remains an open question whether these improved architectures are *equally effective* in 3D point cloud segmentation, especially considering the internal differences between 3D point clouds and images.

To answer this question, we first conduct an in-depth analysis of U-Net and its variants, followed by exploratory experiments to verify whether existing architectures can be smoothly tailored to 3D point clouds. Unfortunately, the U-Net++ architecture, which has been successful in 2D image segmentation, did not yield significant improvements in the segmentation of 3D point clouds (as shown in Fig. 1). This can be attributed to the orderless and irregular nature of 3D point clouds, whereby significant information loss occurs during aggressive upsampling and downsampling, especially considering most existing approaches usually adopt primitive downsampling (random sampling or farthest point sampling), and upsampling scheme (nearest interpolation, and trilinear interpolation). Moreover, within the overarching framework of the network architecture, the successive reduction (high-to-low) and enlargement (low-to-high) of feature dimensions can induce significant noise that accumulates over layers and has a profound impact on segmentation accuracy. In this case, simply aggregating features at varying scales may not be conducive to segmentation accuracy, but increasing the difficulty of optimization due to the sizeable differences in features between different levels.

Bearing this in mind, we hereby propose a general, conceptually simple, yet highly effective point cloud semantic segmentation architecture, termed U-Next. The key of our U-Next is to leverage the most basic U-Net L^1 sub-network, as shown in Fig. 2-(B), which only performs the downsampling and upsampling operations once, thus has the minimal semantic gap and no extra noise, as the building blocks to build the U-Next architecture. That is, stacking as many U-Net L^1 codecs as possible to learn point features from different scales, and enabling local feature maps free to flow laterally, upward, or downward, as shown in Fig. 2. The incorporation of additional scales in our framework serves to strengthen the representation of the input data, thereby resulting in an improved overall accuracy. Finally, we apply multi-level deep supervision at each intermediate node, easing the difficulty in learning and optimization of each constituent U-Net L^1 sub-network. Experiments on three public benchmarks, including S3DIS (Armeni et al., 2016), Toronto3D (Tan et al., 2020), and SensatUrban (Hu et al., 2021a), show that our framework can lead to 2.2%~10.1% improvements in mIoU score, based on the widely-used RandLA-Net network (Hu et al., 2020). Moreover, our framework can be seamlessly integrated into existing segmentation networks, such as PointNet++ (Qi et al., 2018), BAAF-Net (Qiu et al., 2021), LACV-Net (Zeng et al., 2024b), and Point Transformer (Hengshuang et al., 2021), resulting in improved segmentation accuracy. In particular, Point Transformer with our U-Next achieved 72.7% performance on S3DIS Area 5 (2.2% improvement). Remarkably, these improvements were achieved without incurring any visible additional computational costs, thus highlighting the versatility and effectiveness of our U-Next framework. Overall, our key contributions can be summarized as:

- We provide an in-depth analysis of U-Net architecture evolution in the field of 3D point clouds, identifying U-Net L^1 sub-network as a suitable component for fine-grained point cloud segmentation.
- We propose a general and effective segmentation architecture U-Next, by stacking multiple fundamental U-Net L^1 codecs with multi-level deep supervision.
- Extensive experiments on three large-scale benchmarks and up to four baseline approaches demonstrate the effectiveness and generality of our U-Next architecture.

2. Related work

2.1. Learning to segment 3D point clouds

To analyze unstructured 3D point clouds, early works usually utilize hand-crafted feature descriptors (Landrieu et al., 2017; Rusu et al., 2009), while recent works are mainly based on data-driven deep neural architectures. Here, we briefly review the related learning-based techniques, including projection-based, voxel-based, and point-based methods.

Considering the orderless and unstructured of point clouds, directly applying standard convolution neural networks on 3D point clouds remains challenging. Consequently, methods of projection-based (Chen et al., 2017a; Lang et al., 2019; Yu et al., 2018) and voxel-based (Le and Duan, 2018; Graham et al., 2018b; Maturana and Scherer, 2015) are designed to first convert unstructured point clouds into intermediate regular 2D grid images or 3D dense voxels through projection or voxelization steps, and then leverage the successful CNNs to learn from the structured representation. Albeit promising results, the principal limitation of these techniques resides in information loss induced by projection or voxelization steps, especially for learning fine-grained features for 3D point cloud scenes. In contrast, point-based methods are designed to directly operate on irregular point clouds without using intermediate representations. The pioneering PointNet (Qi et al., 2017) learns point-wise feature representations by leveraging the shared Multi-Layer Perceptrons (MLPs). Later, there has been a large amount of point-based approach introduced to learn from unstructured point clouds (Qi et al., 2018; Thomas et al., 2019; Hu et al., 2020; Qiu et al., 2021; Yan et al., 2020; Fan et al., 2021; Shuai et al., 2021; Zeng et al., 2022a). Several representative research works have incorporated

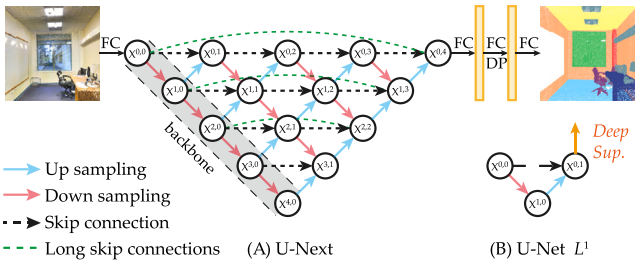


Fig. 2. Illustration of the proposed U-Next and U-Net L^1. (A) The detailed architecture of the proposed U-Next. (B) The U-Net L^1 sub-network with deep supervision. FC: Fully Connected layer; DP: Dropout; Deep Sup.: Deep Supervision.

Recurrent Neural Networks (RNNs) (Ye et al., 2018; Huang et al., 2018), Graph Neural Networks (GNNs) (Landrieu and Simonovsky, 2018; Wang et al., 2019a; Li et al., 2019; Zeng et al., 2022b), and Transformer architectures (Guo et al., 2021; Hengshuang et al., 2021; Yan et al., 2020), to achieve better performance on different tasks.

2.2. Multi-scale feature fusion

Multi-scale fusion has been widely studied and successfully applied in 2D image processing (Cai et al., 2016; Chen et al., 2017b; Samy et al., 2018; Saxena and Verbeek, 2016; Sun et al., 2019). Existing works are mainly built upon the encoder–decoder framework, with different-scale or different-resolution feature maps in the encoder or decoder. Summation or concatenation operations are usually used to fuse different feature maps (Newell et al., 2016; Badrinarayanan et al., 2017; Zhao et al., 2017; Wang et al., 2016). To effectively analyze point clouds which are composed of varying-size objects, a handful of research works also started to explore multi-scale feature fusion in 3D point clouds. The pioneering PointNet++ (Qi et al., 2018) explores multi-scale grouping and abstraction to aggregate features from different scales. BAAF-Net (Qiu et al., 2021) introduce an adaptive fusion strategy that integrate fine-grained representations extracted from multi-resolution feature maps adaptively.

2.3. Learning with deep supervision

Deep supervision was first proposed to tackle the issues of gradient vanishing and slow convergence speed of training deep neural networks (Lee et al., 2015; Szegedy et al., 2015). As an effective training technique, it has been also introduced to different tasks to improve performance (Zhu et al., 2017; Chen et al., 2017c; Tompson et al., 2015; Sun et al., 2019). Lee et al. (2015) illustrated that the incorporation of deep supervision improves the learning capacity. This improvement enforces intermediate layers to capture discriminative features, promoting rapid convergence and facilitating network regularization. Dou et al. (2016) proposed the paradigm to deal with the optimization issues during network training. This deep supervision paradigm involves combining predictions derived from diverse resolutions of feature maps. Deep supervision can also employ to U-Net-like architecture, where the intermediate feature maps are treated as the output of sub-networks. U-Net++ (Zhou et al., 2019) designs a deep supervision scheme with a redesigned skip connection in U-Net architecture to further improve the quality of feature maps. U-Net3+ (Huang et al., 2020) achieves deep supervision by learning from aggregated feature maps hierarchically.

2.4. Improving U-Net architecture

The original U-Net (Ronneberger et al., 2015), which is proposed with symmetrical encoder–decoder architecture, is widely used in image segmentation. In particular, skip connections are used to combine the high-level semantic features from the decoder and low-level details from the encoder. Based on this architecture, a number of improved network architectures have been further proposed. Res-UNet (Xiao et al., 2018) and Dense-UNet (Guan et al., 2019) replace each sub-module of the U-Net with a residual connection or a dense connection, respectively. U-Net++ (Zhou et al., 2019) redesigns skip connection pathways to explore the optimal depth by fusing multiple U-Net sub-networks with varying depths, where each sub-network is connected through a series of nested, dense skip pathways. U-Net3+ (Huang et al., 2020) proposes full-scale skip connections to fully exploit multi-scale feature representations. TransUNet (Chen et al., 2021) introduces Transformer into U-Net to capture the best of both models. This method not only encodes strong global context, but leverages low-level CNN features.

Similarly, this paper also aims to fundamentally improve the U-Net framework for 3D point cloud segmentation. However, in contrast to existing techniques, we first systematically analyze the inherent drawbacks of existing architectures, and then propose our U-Next architecture by stacking a maximum number of sub-networks with minimal codec semantic gap, to effectively learn from non-uniform and unstructured 3D point clouds.

3. The proposed U-Next

In this section, we proposed a novel general segmentation architecture U-Next. First, we revisit the limitations of the U-Net-Like architecture and conduct an in-depth analysis of its challenges when applied in the 3D domain (Section 3.1&3.2). Second, we proposed U-Next, an architecture that is entirely comprised of U-Net L^1. This architecture is designed to optimize the stacking of codec sub-networks and fuse multi-scale feature maps efficiently, thereby enabling the reconstruction of high-resolution representations (Section 3.3&3.4). Third, we design multi-level deep supervision for each U-Net L^1 sub-network to improve the learning ability of local feature maps with multi-scales (Section 3.5). Finally, we introduce the implement details of U-Next (Section 3.6).

3.1. Revisit U-Net-like architectures

U-Net-like architectures have been widely used in modern segmentation models. As shown in Fig. 3(A), this network is based on a symmetrical encoder–decoder architecture, with skip connections playing a crucial role in integrating the deep, semantically-rich feature maps from the decoder sub-network with the shallow, fine-grained feature maps from the encoder sub-network. Despite achieving promising results in the recovery of fine-grained details for dense prediction tasks like semantic and instance segmentation, this architecture has limitations. Specifically, it only performs feature aggregation at the same scale, which is restrictive and lacks a solid theoretical foundation. Furthermore, the combined feature maps have been shown to be semantically dissimilar (Zhou et al., 2019).

Challenges. Recently, a handful of advanced architectures such as U-Net+ (Fig. 3(B)) and U-Net++ (Zhou et al., 2019) (Fig. 3(C)) have been proposed. These architectures assemble various levels of U-Net as sub-networks with deep supervision to effectively leverage multi-scale feature maps, and make efforts in reducing semantic gaps between combined feature maps in medical image segmentation. Nevertheless, the application of these architectures to 3D point clouds remains a highly challenging task. **First**, most existing architectures can be fundamentally viewed as an ensemble of U-Nets with varying different depths. That is, the semantic gap between the combined features in deeper U-Nets remains substantial, potentially hindering network learning and optimization. **Second**, Considering aggressive upsampling and downsampling are frequently used in point cloud neural architectures, coupled with the orderless and unstructured nature of 3D point clouds, the recovered full-resolution feature maps from horizontal and low-to-high features fusion may still be inadequate for fine-grained semantic segmentation.

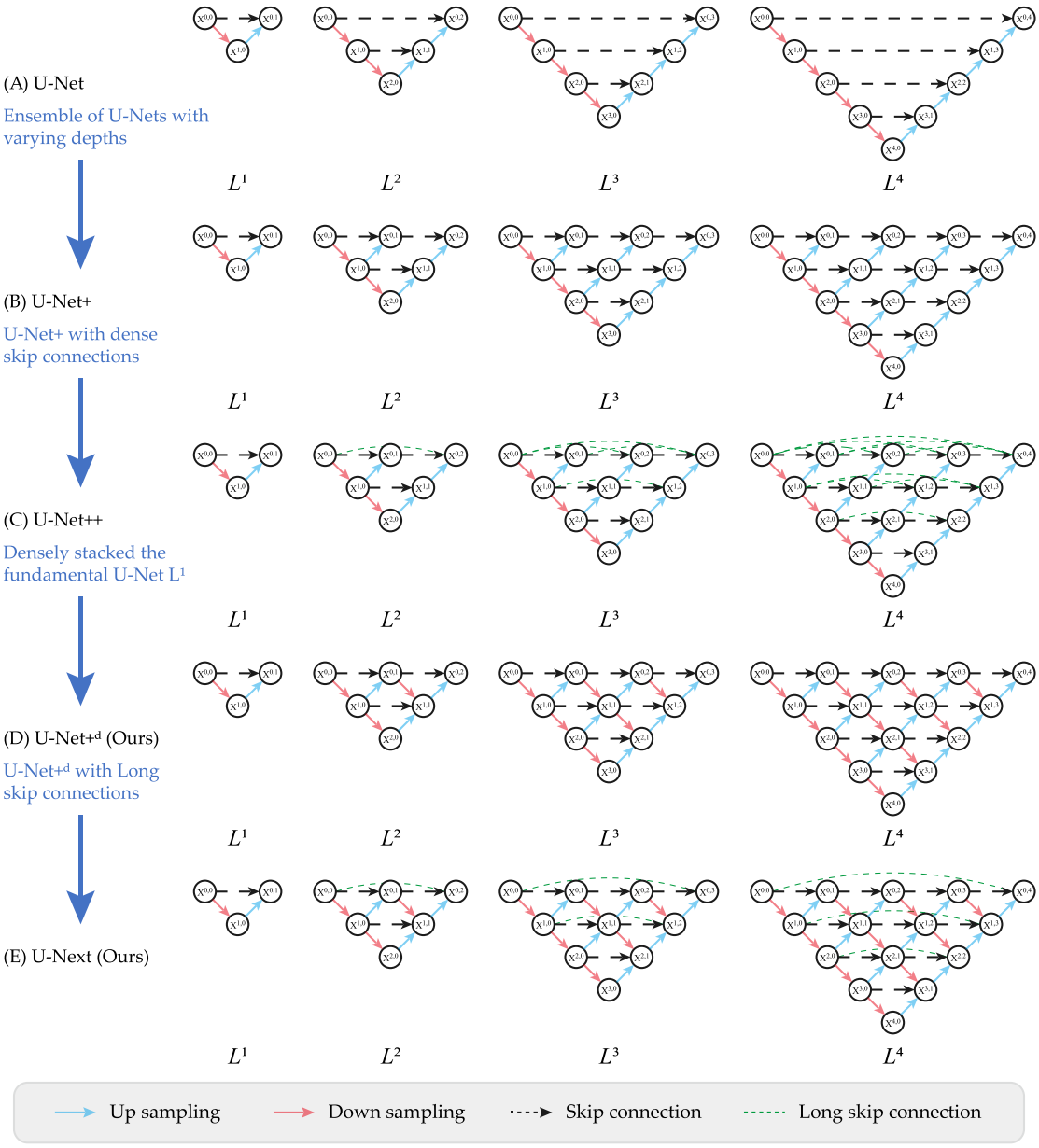


Fig. 3. Illustration of the evolution from U-Net to U-Next. The key contributions of each architecture are identified. (A) is the original U-Net (Ronneberger et al., 2015), (B) and (C) are U-Net+ and U-Net++ proposed by Zhou et al. (2019), and (E) and (F) are the proposed U-Net+^{d} and U-Next. It is clear that U-Net+ and U-Net++ are fundamentally an ensemble of different levels of U-Net with shared encoding layers. Similarly, the proposed U-Next can also be seen as an integration of U-Net, but entirely consists of U-Net L^1 sub-networks, which means each node in U-Next incorporates horizontal, low-to-high and high-to-low features, while also with minimal semantic gaps. Additionally, different from U-Net++, our U-Next does not use dense skip connections, primarily because of the efficiency consideration.

3.2. Motivations and solutions

Motivations. As shown in Fig. 4, we further provide an intuitive comparison of architecture and operation between 2D and 3D networks. Clearly, the input 2D images are structured and regular, while 3D point clouds are discrete and irregular. As such, although aiming for the same segmentation task, the following differences exist: **(1) Sampling**. Max-pooling with strides is usually used to reduce the resolution in 2D images, while random or farthest point sampling are usually used to down-sample the 3D point clouds; **(2) Neighborhood Definition**. Considering the non-uniform distribution and uneven density of 3D point clouds, KNN and spherical search are usually used to determine the neighborhood, while the neighborhood can be easily determined by the neighborhood pixels in 2D images; **(3) Low-level operations**. Standard CNN is used as the fundamental operation for 2D segmentation, while shared MLPs are usually taken as the basic operation in 3D point clouds. Overall, considering the orderless and unstructured nature of 3D point clouds, coupled with unstable neighborhoods and aggressive down-sampling operations, it should be more careful to aggregate the multi-scale features from different nodes in U-Net architecture. Motivated by this, we go back to the beginning and stack as many U-Net L^1 codecs as possible in our U-Next framework, and introduce multi-level deep supervision, to ensure that the aggregated features have minimal semantic gaps, further leading to increasing performance in semantic segmentation task.

Solutions. Given the inevitable information loss during downsampling and upsampling, along with the growing semantic gap between codecs with increasing U-Net depth, we were motivated to re-examine the U-Net architecture and propose a dedicated framework for point cloud semantic segmentation. Drawing inspiration from the concept

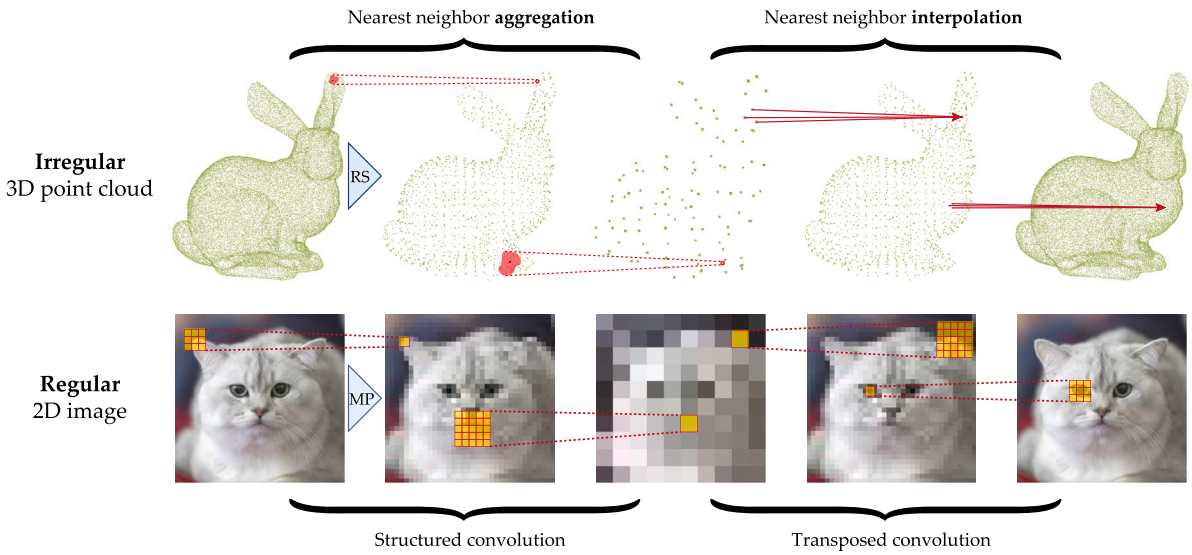


Fig. 4. Illustration of irregular 3D point cloud processing and regular 2D image processing. RS: random sampling; MP: max pooling. Typically, max pooling is used to down-sample regular 2D images, in this process, structured CNNs are used to capture local feature maps; random sampling or farthest point sampling is used to down-sample irregular 3D point clouds, in this process, unstructured nearest neighbor aggregation are used to capture local feature maps. Then, structured transposed CNNs and pixel shuffle are usually used for 2D images to recover high-resolution information, whereas uneven nearest neighbor interpolations are usually used for 3D point clouds.

of building blocks, we wondered if it was feasible to build U-Net architecture from the fundamental U-Net L^1 sub-network, since there is a single downsampling and upsampling operation in U-Net L^1 sub-network, resulting in a naturally small semantic gap. Next, considering the shallow structure and limited learning capacity of the U-Net L^1 sub-network, we adopt a straightforward approach to enhance the feature learning capacity by densely nesting and stacking multiple U-Net L^1 codecs. This allows the network to learn multi-scale features, ultimately leading to improved performance. Additionally, to ensure smooth gradient backpropagation and reduce optimization difficulty during training, it is desirable to further introduce a dedicated deep supervision scheme. To summarize, a feasible solution is to take the fundamental U-Net L^1 codecs as building blocks for our framework, which can simultaneously minimize the semantic gap and recover fine-grained feature maps, thus enabling high-quality 3D semantic segmentation.

3.3. U-Next: Building backbone architecture with U-Net L^1 sub-networks

Before building the final U-Next framework, we first proposed a hierarchically ensemble architecture based on the fundamental U-Net L^1 codecs, as shown in Fig. 3 (D). It can be seen that this architecture explicitly adds the connections between different-level decoder nodes, compared with the previous U-Net+ architecture, enabling each node receives information from other parallel and adjacent (lower and higher) nodes, leading to sufficient and comprehensive information fusion. This architecture can be also viewed as progressively stacking as many U-Net L^1 codec as possible, which naturally has the minimal codec semantic gap. Further, we enriched U-Net+^{d} architecture with long skip connections, to build the final U-Next architecture, as shown in Fig. 3 (E). Different from the nested, dense skip pathways used in U-Net++ framework, our U-Next architecture only adds simple sparse skip connections similar to the original U-Net architecture.

(1) Building U-Net+^{d} from U-Net L^1

As shown in Fig. 3(E), each node in U-Net+^{d} receives information from other parallel and adjacent (lower and higher) nodes, leading to comprehensive and sufficient multi-scale information fusion. Specifically, We use $x^{i,j}$ to denote the output of node $X^{i,j}$ (each node in the same horizontal row with the same i, and in the same slash along the down-sampling direction with the same j). The stack of feature maps in U-Net+^{d} represented by $x^{i,j}$ can be formulated as follows:

$$x^{i,j} = \begin{cases} C[x^{i-1,j}], & j = 0 \\ C[x^{i-1,j}, \mathcal{U}(x^{i-1,j+1})], & j > 0, i = 0 \\ C[x^{i-1,j}, \mathcal{U}(x^{i-1,\ j+1}), \mathcal{D}(x^{i-1,j-1})], & j > 0, i > 0 \end{cases} \tag{1}$$

where $C(\cdot)$ is coding block (usually a feature extraction module), $\mathcal{D}(\cdot)$ and $\mathcal{U}(\cdot)$ are the down-sampling operation and up-sampling operation, and $[\cdot]$ is the concatenation operation.

(2) Building U-Next from U-Net+^{d}

We build U-Next architecture based on U-Net+^{d} by introducing long skip connections. As shown in Fig. 3(F), the earlier feature maps are travel through the long skip connections of U-Next to integrate with the later feature maps. Specifically, concatenating feature maps with additional long skip connections are computed by fusing the first node in ith row with the node $X^{i,j}$, if and only if node $X^{i,j}$ is the last node in row i, as follows:

$$x^{i,j} = \begin{cases} C[x^{i-1,j}, \mathcal{U}(x^{i-1,j+1}), x^{i,0}], & i = 0 \\ C[x^{i-1,j}, \mathcal{U}(x^{i-1,\ j+1}), \mathcal{D}(x^{i-1,j-1}), x^{i,0}], & i > 0 \end{cases} \tag{2}$$

3.4. Node connectivity pipeline

To further illustrate how we aggregate features from different nodes, we use pseudo-code to show the node connectivity pipeline in the U-Next architecture, as shown in Algorithm 1. Basically, as shown in Fig. 5 (A-C), the node with $j = 0$ receives only the output of previous coding layer as input, the node with $i = 0$ ($j > 0$) receives two inputs, both from the same U-Net L^1 sub-network, and the node with $i > 0$ and $j > 0$ receives the output of previous coding layer as input, and also the two inputs from the same U-Net L^1 sub-network. U-Next introduces additional long skip connections on top of U-Net+^{d}, instead of the dense skip connections of U-Net++, considering that there is no apparent performance improvement (validated in Section 4.3.4). Fig. 5 (D-E) further clarifies the stack of feature maps in additional long skip connections represented by $x^{i,j}$, if and only if node $X^{i,j}$ is the last node in ith row.

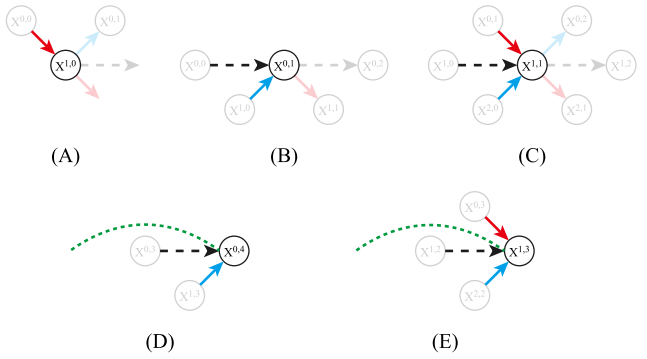


Fig. 5. Node connectivity pipeline for different types of nodes in U-Net+^{d} and U-Next. The input of each node type is prominently identified. U-Net+^{d} is only represented in (A-C), and U-Next builds on the U-Net+^{d} also includes (D-E).

Algorithm 1: Node Connectivity Pipeline

input : Initial embedding
output: Recovered ori-resolution feature map

```
for j in range(N) do
    if j==0 then
        for i in range(N − j) do
            x = Coding(x);
            List [j] ← x;
            x = DownSampling(x);
        end
    end
    else if j>0 then
        for i in range(N − j) do
            x0 = List [j − 1][i];
            x1 = UpSampling(List [j − 1][i + 1]);
            if i>0 then
                x2 = DownSampling(x);
            end
            if i+j==N-1 then
                // The last Node of each layer
                x3 = List [0][i];
            end
            x = concatenate all the above;
            x = Coding(x);
            List [j] ← x;
        end
    end
    // switch to next Sequence ;
end
```

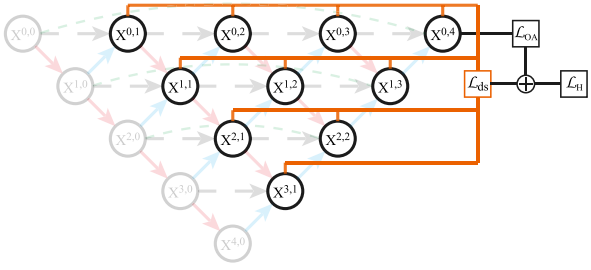


Fig. 6. Illustration of deep supervision. All the sub-networks are supervised. $\mathcal{L}_{ds}$: loss of the multi-level deep supervision; $\mathcal{L}_{OA}$: loss of the overall network; $\mathcal{L}_H$: hybrid loss.

3.5. Learning with multi-level deep supervision

Up to here, multiple U-Net L^1 sub-networks have been hierarchically stacked in our framework. To further learn hierarchical representations from the aggregated feature maps, we introduce multi-level deep supervision into our framework from the perspective of optimization and learning. As shown in Fig. 6, different from U-Net++ only apply deep supervision on the generated full-resolution feature maps, our U-Next framework have explicitly applied deep supervision to each decoder node at different levels (*i.e.,* for each U-Net L^1 sub-network). By introducing ground-truth supervision signal at different decoding stages of our framework, smooth gradient propagation and easy optimization of the network are more likely to be achieved. Therefore, all the constituent U-Net L^1 sub-network are trained simultaneously to improve the overall learning capacity of multi-scale feature maps.

(1) Loss of multi-level deep supervision

To achieve multi-level deep supervision, the output of decoder in *each* U-Net L^1 sub-network is fed into a plain 1×1 convolution layer to align with ground truth. After that, we follow the widely-used weighted cross-entropy loss to calculate the loss of each U-Net L^1 sub-network. Finally, the total loss of deep supervision is defined as the average of all the losses of the U-Net L^1 sub-networks, as follows:

$$\mathcal{L}_{ds}(Y,P) = -\frac{1}{N}\sum_{i=0}^{L}\sum_{j=0}^{L-i}\sum_{c=0}^{C} y_c^i \log p_c^{i,j} \tag{3}$$

where y_c^i is the true labels of the ith row, $p_c^{i,j}$ is the predicted labels of node $X^{i,j}$, C is the number of label categories, L is the number of architecture levels, and N is the number of decoder nodes. In the multi-level deep supervision, all the sub-networks are trained, enhancing the learning ability of each sub-network for local features.

(2) Hybrid loss

As illustrated in Fig. 6, after the final node ($X^{0,4}$), the fully-connected layers are used to predict semantic labels of all input points. Therefore, we also employ the weighted cross-entropy loss to calculate the loss of overall network by the final prediction results. Further, the final loss function of the framework is a hybrid loss which combined the loss of the multi-level deep supervision and overall network, as follows:

$$\mathcal{L}_H = \mathcal{L}_{ds} + \mathcal{L}_{oa} \tag{4}$$

3.6. Details of the network architecture

As illustrated in Fig. 2, we present the general architecture of U-Next. The architecture gradually reduces the resolutions of point clouds from: ($\frac{N}{4} \rightarrow \frac{N}{16} \rightarrow \frac{N}{64} \rightarrow \frac{N}{256} \rightarrow \frac{N}{512}$), whereas the channel dimension of the features increases as: ($16 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$). Note that the number of points and feature dimensions are the output of each coding block, and in the same layer (same i), the output of each coding block has the same points and same dimension of features. In the process, point clouds with different resolutions are generated, and the proposed U-Next utilizes multiple sub-networks to fuse the multi-scale feature maps. Finally, we use three fully-connected layers ($64 \rightarrow 32 \rightarrow c$, the drop-out ratio is set as 0.5), and a softmax layer to transform the abstract semantic features to classification scores for each point in the point cloud scene (see Table 1).

4. Experiments

4.1. Experiment setup

To comprehensively evaluate the effectiveness of our U-Next, experiments are conducted on widely-used indoor S3DIS dataset (Armeni et al., 2016), outdoor LiDAR-based Toronto3D dataset (Tan et al., 2020), and urban-scale photogrammetric SensatUrban dataset (Hu

Table 1
The output of points and feature dimensions of coding blocks in our U-Next. The output size of each node in the same layer is the same.

Layer	Points	Dimension
L_0	$N/4$	16
L_1	$N/16$	64
L_2	$N/64$	128
L_3	$N/256$	256
L_4	$N/512$	512

et al., 2021a). To quantitatively evaluate the segmentation performance, we take the per-class Intersection over Union (IoUs), mean IoU (mIoU), and Overall Accuracy (OA) as evaluation metrics:

$$OA = \frac{\sum_{i=1}^{n} TP_i}{N} \tag{5}$$

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{6}$$

$$mIoU = \frac{\sum_{i=1}^{n} IoU_i}{n} \tag{7}$$

where TP represents true positive samples, FP represents false positive samples, FN represents false negative samples, i represents the i_{th} semantic class, n represents total semantic classes and N represents total points.

Implementation Details. Given the simplicity and high efficiency of the point-based RandLA-Net (Hu et al., 2020), we have selected it as the baseline model to evaluate the effectiveness of our U-Next architecture. It is important to note that our architecture is not constrained to RandLA-Net, more baseline models are demonstrated in Section 4.3. Specifically, the local feature aggregation module used in RandLA-Net is integrated as coding blocks for U-Next as shown in Eq. (1) & 2. During the network training, we input 40,960 points into U-Next, the batch size is set to 4 with maximum 100 epochs. Adam optimizer (Kingma and Ba, 2015) is employed with default parameters to optimize the hybrid loss function. We follow RandLA-Net (Hu et al., 2020) to set K=16 for KNN, and the initial learning rate is set to 0.01. All experiments are conducted on an NVIDIA GeForce RTX3090 GPU. We utilize simple random down-sampling to improve computational efficiency during training rather than high cost furthest point sampling.

4.2. Semantic segmentation results

4.2.1. Evaluation on S3DIS

The Stanford Large-Scale 3D Indoor Spaces (S3DIS) (Armeni et al., 2016) was acquired from six extensive indoor environments, encompassing a total of 272 rooms. Points in this dataset are categorized into 13 semantic classes, and each point is characterized by a 6D vector (comprising coordinates xyz and color rgb). To comprehensively evaluate segmentation performance, we withhold Area 5 during training and use it for testing, and also follow the widely-used 6-fold cross-validation strategy.

Table 2 shows quantitative results of our U-Next framework with RandLA-Net (Hu et al., 2020) and Point Transformer (Hengshuang et al., 2021) as baselines on S3DIS Area 5. From the result, we can see that by replacing the original U-Net (Ronneberger et al., 2015) with our U-Next, the overall segmentation performance and all the categories can be significantly improved. We also noticed that our U-Next with Point Transformer almost achieves the top performance in mIoU score. Subsequently, we report the quantitative results achieved by the proposed method with RandLA-Net and Point Transformer (Hengshuang et al., 2021) as baselines, and other baselines on the S3DIS six-fold in Table 3. It can be seen that RandLA-Net and Point Transformer equipped with the proposed U-Next architecture achieves strong segmentation performance with mIoU score of 73.2% and 75.7%, outperforming the baselines by 3.2% and 2.2%. It is also noted that the proposed architecture consistently improves the segmentation performance of categories. In particular, the performance of our method is superior to the baseline in large plane categories such as *walls*, and small instance categories such as *boards*, *columns*, and *beams*. This clearly shows that stacking more sub-networks to capture local geometric details can effectively improve the segmentation performance.

Fig. 7 provides a qualitative visual comparison o S3DIS (Armeni et al., 2016). The results clearly demonstrate that our segmentation of boards, columns, doors, and bookcases outperforms RandLA-Net in terms of visual quality. Furthermore, the proposed U-Next exhibits a capacity to segment object boundaries with enhanced smoothness and accuracy. This validates that repeated fusion of information from sub-networks with minimal semantic gaps helps to rebuild fine-grained perception to alleviate irregular and non-uniform sampling and aggregation in 3D point clouds.

4.2.2. Evaluation on Toronto3D

The Toronto3D dataset (Tan et al., 2020) was obtained covering around 1 km of roadway scenarios in Toronto, Canada. Toronto3D is partitioned into four tiles, each covering around 250 m, and points are labeled into 8 semantic categories. During training, both 3D coordinates and color attributes are available to use. Following (Tan et al., 2020), Area 2 is designated as the test set, while the remaining three areas are used as training sets.

Table 4 presents the quantitative results attained by the U-Next and other state-of-the-art method on the Toronto3D. Considering several baselines do not use color information as input on this dataset, we thereby report the performance of our method with or without utilizing rgb information, for a fair comparison. It can be seen that the proposed RandLA-Net (U-Next) achieves the best performance with/without color attributes. In particular, the mIoU scores of the proposed method improve by 1.5% and 2.2% respectively, compared with the baseline RandLA-Net. It is evident that our U-Next framework can steadily improve the segmentation performance of RandLA-net in all of the classes. In particular, the performance of our method is superior to the baseline in classes with vague local geometric information such as road markings and poles. This is primarily because our framework stacks more sub-networks to capture sufficient local details.

Fig. 8 qualitatively shows the visual comparison for Toronto3D (Tan et al., 2020) using rgb as input. The segmentation results illustrate a pronounced superiority in our U-Next for car, nature, and fence segmentation compared to RandLA-Net. Notably, in the case of crosswalk semantic segmentation, U-Next demonstrates heightened accuracy and clearer delineation of boundaries.

4.2.3. Evaluation on SensatUrban

The SensatUrban dataset (Hu et al., 2021a) is a photogrammetric point cloud dataset that spans across 7.6 km^2 of the urban landscape, comprising almost 3 billion points that are densely annotated with 13 semantic categories. We follow the official splits to train our network and evaluate the performance on the online test server.

Table 5 presents the quantitative results obtained by the proposed U-Next and other state-of-the-art methods on the SensatUrban dataset. We can see that the proposed method significantly improves the mIoU score from 52.7% to 62.8%, with clear improvements in almost all categories. We also notice that the proposed method can achieve an IoU score of 36.8% on the challenging *railway* class, primarily because the key of our method is to stack as many U-Net L^1 codecs as possible with multi-level deep supervision, hence less information loss and small semantic gap are achieved, further leading to improved segmentation performance, especially for small objects. This can be also verified in the categories of class wall and footpath.

Table 2
Quantitative comparisons with the state-of-the-art methods on S3DIS (*Area 5*). Bold indicates the best result.

Method	OA	mIoU	Ceil.	Floor	Wall	Beam	col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
KPConv (Thomas et al., 2019)	–	67.1	92.8	97.3	82.4	0.0	23.9	58.0	69.0	81.5	91.0	75.4	75.3	66.7	58.9
PointASNL (Yan et al., 2020)	87.7	62.6	94.3	98.4	79.1	0.0	26.7	55.2	66.2	83.3	86.8	47.6	68.3	56.4	52.1
DenseKPNet (Li et al., 2022)	90.8	68.9	94.8	98.6	84.6	0.0	25.1	61.0	76.8	81.7	92.0	77.1	69.8	72.0	61.6
LGGCM (Du et al., 2022)	88.8	63.3	94.7	98.3	81.5	0.0	35.9	63.3	43.5	88.4	80.2	68.8	55.7	64.6	47.8
LACV-Net (Zeng et al., 2024b)	89.3	65.9	92.6	97.1	81.8	0.0	36.5	64.1	46.7	77.5	88.2	76.6	71.0	68.1	56.1
FA-ResNet (Zhan et al., 2023)	89.0	68.1	94.0	97.9	82.6	0.0	39.1	64.1	65.2	78.5	87.5	69.8	77.6	73.3	56.0
AGConv (Zhou et al., 2023)	90.0	67.9	93.9	98.4	82.2	0.0	23.9	59.1	71.3	**91.5**	81.2	75.5	74.9	72.1	58.6
SAKS (Chen et al., 2023)	90.8	68.8	**95.2**	98.6	84.1	0.0	27.5	58.5	**75.1**	90.8	81.8	77.0	69.0	73.5	**62.1**
GSLCN (Liang et al., 2023)	90.5	68.1	94.3	98.5	82.9	0.0	20.6	59.4	69.8	83.1	91.4	76.9	75.4	72.5	60.7
MKConv (Woo et al., 2023)	89.6	67.7	92.4	98.2	83.9	0.0	28.5	64.5	65.7	82.4	89.7	73.9	67.5	77.3	55.9
SFL-Net (Li et al., 2023b)	90.3	69.2	94.3	98.4	84.2	0.0	28.3	58.9	73.2	82.9	92.2	76.6	82.2	68.6	59.9
PointNeXt (Qian et al., 2022)	90.6	70.5	94.2	98.5	84.4	0.0	37.7	59.3	74.0	83.1	91.6	77.4	77.2	78.8	60.6
TCFAP-Net (Zhang et al., 2024)	88.9	66.1	93.4	97.8	83.1	0.0	31.2	64.7	45.4	79.9	89.2	74.1	72.5	76.8	57.6
DG-Net (Liu et al., 2024)	90.2	65.9	93.1	97.8	82.6	0.0	33.1	62.2	52.3	80.4	87.5	72.3	70.7	69.8	54.5
RandLA-Net (Hu et al., 2020)	87.2	62.5	92.1	97.3	80.9	0.0	21.4	61.4	37.4	78.3	87.1	65.8	70.4	67.7	52.2
+ U-Next	89.5 +2.3	68.3 +5.8	92.2	97.9	83.9	0.0	30.2	63.1	60.5	82.7	88.9	82.6	75.3	73.6	56.5
Point Transformer (Hengshuang et al., 2021)	90.8	70.4	94.0	98.5	86.3	0.0	38.0	63.4	74.3	89.1	82.4	74.3	80.2	76.0	59.3
+ U-Next	**91.2** +0.4	**72.7** +2.2	94.4	**98.7**	**86.5**	0.0	**40.4**	**64.9**	75.0	90.9	**92.5**	**83.1**	**81.8**	77.2	60.3

Table 3
Quantitative comparisons with the state-of-the-art methods on S3DIS (six-fold cross-validation). Bold indicates the best result.

Method	OA	mIoU	Ceil.	Floor	Wall	Beam	Col.	Wind.	Door	Table	Chair	Sofa	Book.	Board	Clut.
PointNet (Qi et al., 2017)	78.6	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
DGCNN (Wang et al., 2019b)	84.5	55.5	93.2	95.9	72.8	54.6	32.2	56.2	50.7	62.8	63.4	22.7	38.2	32.5	46.8
SPGraph (Landrieu and Simonovsky, 2018)	86.4	62.1	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
DeepGCNs (Li et al., 2019)	85.9	60.0	93.1	95.3	78.2	33.9	37.4	56.1	68.2	64.9	61.0	34.6	51.5	51.1	54.4
KPConv (Thomas et al., 2019)	–	70.6	93.6	92.4	83.1	63.9	54.3	66.1	76.6	57.8	64.0	69.3	**74.9**	61.3	60.3
SC-CNN (Wang et al., 2022b)	–	67.4	93.6	94.1	83.1	50.5	33.0	60.3	69.6	63.0	72.3	63.8	64.3	60.6	57.1
DenseKPNet (Li et al., 2022)	89.3	71.9	94.5	95.0	84.3	63.4	57.4	67.2	77.2	60.5	71.0	70.7	70.7	62.6	60.7
BAAF-Net (Qiu et al., 2021)	88.9	72.2	93.3	96.8	81.6	61.9	49.5	65.4	73.3	72.0	83.7	67.5	64.3	67.0	62.4
BAF-LAC (Shuai et al., 2021)	88.2	71.7	92.5	95.9	81.3	63.2	57.8	63.0	**79.9**	70.3	74.6	60.6	67.2	65.3	60.4
SCF-Net (Fan et al., 2021)	88.4	71.6	93.3	96.4	80.9	64.9	47.4	64.5	70.1	71.4	81.6	67.2	64.4	67.5	60.9
LEARD-Net (Zeng et al., 2022a)	89.1	72.5	94.2	96.9	81.8	65.1	50.9	69.9	72.5	70.6	78.2	68.6	67.2	66.1	60.3
LACV-Net (Zeng et al., 2024b)	89.7	72.7	94.5	96.7	82.1	65.2	48.6	69.3	71.2	72.7	78.1	67.3	67.2	70.9	61.6
CBL (Tang et al., 2022)	89.6	73.1	94.1	94.2	85.5	50.4	58.8	**70.3**	78.3	**75.7**	75.0	71.8	74.0	60.0	62.4
HADF-Net (Zhou et al., 2024)	89.2	72.7	93.5	97.0	82.3	62.2	51.2	65.8	71.9	72.7	82.0	67.7	65.8	71.4	63.1
RandLA-Net (Hu et al., 2020)	88.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
+ U-Next	89.5 +1.5	73.2 +3.2	93.6	96.9	84.2	**66.1**	54.6	67.6	75.5	73.6	74.5	62.9	66.2	**74.0**	61.7
Point Transformer (Hengshuang et al., 2021)	90.2	73.5	94.3	97.5	84.7	55.6	58.1	66.1	78.2	77.6	74.1	67.3	71.2	65.7	64.8
+ U-Next	**91.0** +0.8	**75.7** +2.2	**94.7**	**97.9**	**85.7**	59.4	**59.9**	68.0	78.3	74.9	84.0	**74.9**	69.3	71.2	**66.7**

Table 4
Quantitative comparisons with the state-of-the-art methods on Toronto3D (*Area 2*). Bold indicates the best result.

RGB	Method	OA	mIoU	Road	Road mrk.	Nature	Buil.	Util. line	Pole	Car	Fence
No	PointNet++ (Qi et al., 2018)	92.6	59.5	92.9	0.0	86.1	82.2	60.9	62.8	76.4	14.4
	DGCNN (Wang et al., 2019b)	94.2	61.7	93.9	0.0	91.3	80.4	62.4	62.3	88.3	15.8
	MS-PCNN (Ma et al., 2021)	90.0	65.9	93.8	3.8	93.5	82.6	67.8	71.9	91.1	22.5
	KPConv (Thomas et al., 2019)	95.4	69.1	94.6	0.1	96.1	91.5	87.7	**81.6**	85.7	15.7
	TGNet (Li et al., 2020)	94.1	61.3	93.5	0.0	90.8	81.6	65.3	62.9	88.7	7.9
	MS-TGNet (Tan et al., 2020)	95.7	70.5	94.4	17.2	95.7	88.8	76.0	73.9	94.2	23.6
	LACV-Net (Zeng et al., 2024b)	95.8	78.5	94.8	42.7	96.7	91.4	**88.2**	79.6	93.9	**40.6**
	MVP-Net (Li et al., 2023a)	96.1	75.1	95.1	22.0	96.5	92.8	88.3	85.0	91.8	29.5
	SFL-Net (Li et al., 2023b)	96.0	78.1	94.2	34.0	96.9	93.8	87.1	85.7	93.5	39.7
	RandLA-Net (Hu et al., 2020)	93.0	77.7	94.6	42.6	96.9	93.0	86.5	78.1	92.9	37.1
	+ U-Next	**96.0** +3.0	**79.2** +1.5	**95.1**	**44.8**	**97.2**	**93.5**	87.6	80.5	**94.3**	40.8
Yes	ResDLPS-Net (Du et al., 2021)	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
	BAF-LAC (Shuai et al., 2021)	95.2	82.2	96.6	64.7	96.4	92.8	86.1	83.9	93.7	43.5
	BAAF-Net (Qiu et al., 2021)	94.2	81.2	96.8	67.3	96.8	92.2	86.8	82.3	93.1	34.0
	LEARD-Net (Zeng et al., 2022a)	97.3	82.2	97.1	65.2	96.9	92.8	87.4	78.6	94.5	**45.2**
	LACV-Net (Zeng et al., 2024b)	97.4	82.7	97.1	66.9	97.3	93.0	87.3	83.4	93.4	43.1
	SFL-Net (Li et al., 2023b)	97.6	81.9	97.7	70.7	95.8	91.7	87.4	78.8	92.3	40.8
	TCFAP-Net (Zhang et al., 2024)	97.0	81.9	97.1	64.8	97.2	94.3	87.9	81.9	93.0	38.6
	DG-Net (Liu et al., 2024)	96.8	82.1	97.1	65.2	97.2	92.6	88.1	84.2	93.6	38.7
	RandLA-Net (Hu et al., 2020)	94.4	81.8	96.7	64.2	96.9	94.2	88.0	77.8	93.4	42.9
	+ U-Next	**97.7** +3.3	**84.0** +2.2	**97.3**	**68.6**	**97.7**	**95.2**	**88.4**	**86.1**	**95.1**	43.2

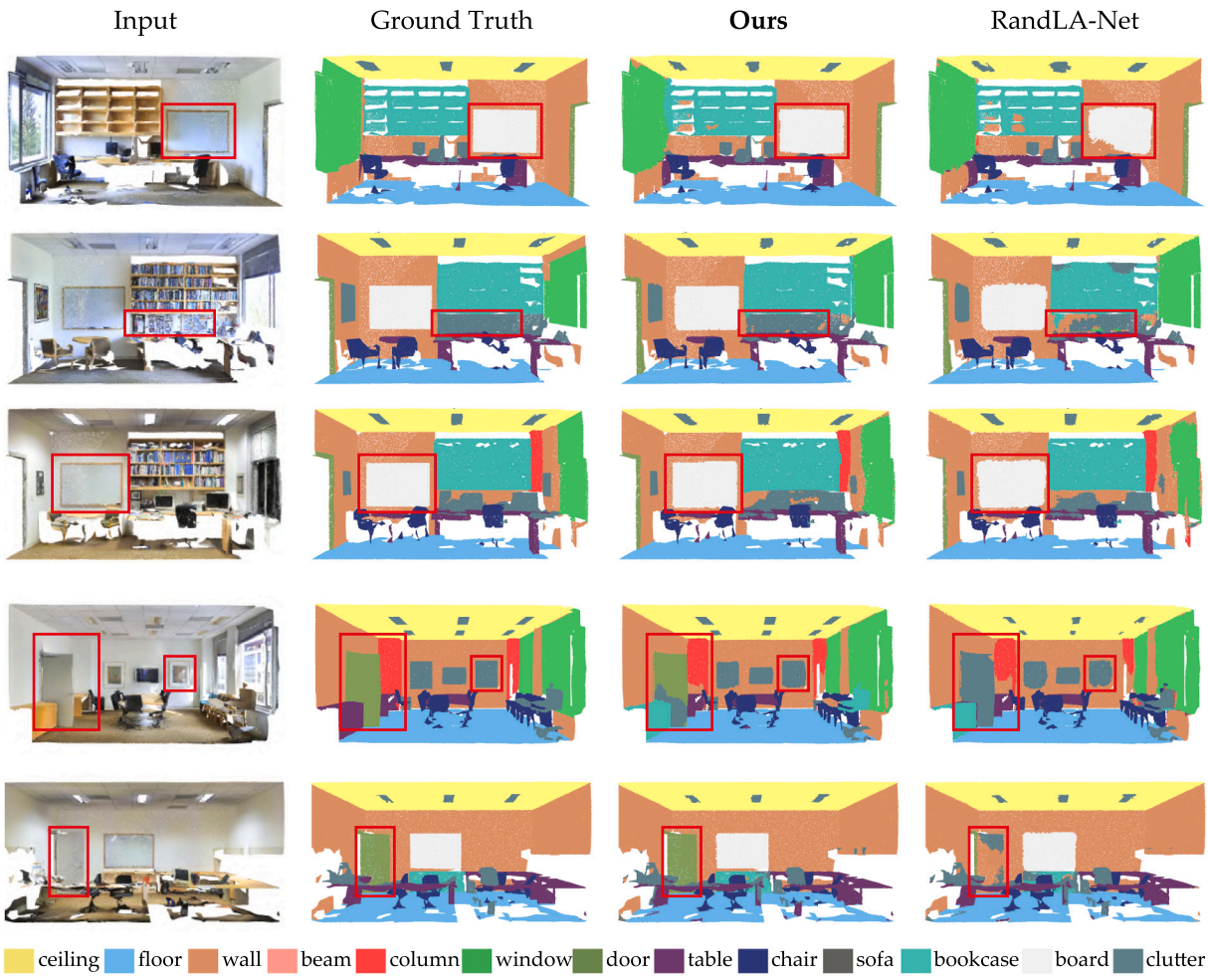


Fig. 7. Visual comparison of semantic segmentation results on S3DIS dataset.

Table 5
Quantitative comparisons with the state-of-the-art methods on SenastUrban. Bold indicates the best result, underline indicates the sub-best result.

Method	OA	mIoU	Ground	Veg.	Buil.	Wall	Bri.	Park.	Rail	Traffic.	Street.	Car	Foot.	Bike	Water
PointNet (Qi et al., 2017)	80.8	23.7	67.9	89.5	80.1	0.0	0.0	3.9	0.0	31.6	0.0	35.1	0.0	0.0	0.0
PointNet++ (Qi et al., 2018)	84.3	32.9	72.5	94.2	84.8	2.7	2.1	25.8	0.0	31.5	11.4	38.8	7.1	0.0	56.9
TagentConv (Tatarchenko et al., 2018)	76.9	33.3	71.5	91.4	75.9	35.2	0.0	45.3	0.0	26.7	19.2	67.6	0.0	0.0	0.0
SPGraph (Landrieu and Simonovsky, 2018)	85.3	37.3	69.9	94.6	88.9	32.8	12.6	15.8	15.5	30.6	22.9	56.4	0.5	0.0	44.2
SparseConv (Graham et al., 2018a)	88.7	42.7	74.1	97.9	94.2	63.3	7.5	24.2	0.0	30.1	34.0	74.4	0.0	0.0	54.8
KPConv (Thomas et al., 2019)	93.2	57.6	87.1	**98.9**	95.3	**74.4**	28.7	41.4	0.0	55.9	**54.4**	**85.7**	40.4	0.0	**86.3**
BAF-LAC (Shuai et al., 2021)	91.5	54.1	84.4	98.4	94.1	57.2	27.6	42.5	15.0	51.6	39.5	78.1	40.1	0.0	75.2
BAAF-Net (Qiu et al., 2021)	92.0	57.3	84.2	98.3	94.0	55.2	48.9	57.7	20.0	57.3	39.3	79.3	40.7	0.0	70.1
LACV-Net (Zeng et al., 2024b)	**93.2**	61.3	**85.5**	98.4	**95.6**	61.9	**58.6**	**64.0**	28.5	62.8	45.4	81.9	42.4	**4.8**	67.7
EyeNet (Yoo et al., 2023)	93.7	62.3	86.6	98.6	96.2	65.8	59.2	64.8	17.9	64.8	49.8	83.1	46.2	11.1	65.4
Rong et al. Rong and Shen (2023)	93.0	62.5	84.2	97.6	94.7	70.5	26.8	59.7	58.1	61.4	49.7	82.7	41.2	2.5	83.1
MVP-Net (Li et al., 2023a)	93.3	59.4	85.1	98.5	95.9	66.6	57.5	52.7	0.0	61.9	49.7	81.8	43.9	0.0	78.2
RandLA-Net (Hu et al., 2020)	89.8	52.7	80.1	98.1	91.6	48.9	40.6	51.6	0.0	56.7	33.2	80.1	32.6	0.0	71.3
+ U-Next	93.0 + 3.2	**62.8** + 10.1	85.2	98.6	95.0	68.2	53.6	60.4	**36.8**	**64.0**	48.9	84.9	**45.1**	0.0	76.2

In Fig. 9, a qualitative visual comparison on the SensatUrban (Hu et al., 2021a) validation set is presented, as ground truth for the online test sets is not accessible. The segmentation results shows that compared to RandLA-Net, which barely segments the railway, the proposed U-Next adeptly segments the railway from ground. Additionally, noteworthy enhancements are observed in the performance of segmenting large parking areas and small traffic roads.

4.3. Ablation studies

Here, we conduct extensive ablation studies on the proposed U-Next architecture from the following four aspects. In this section, we first show the adaptability of the U-Next architecture across varying baselines (Section 4.3.1), and we then show quantitative and qualitative comparisons of the U-Next architecture with other U-Net-like architectures (Section 4.3.2). Consequently, we delve into the evaluation of the effectiveness of multi-level deep supervision (Section 4.3.3) and

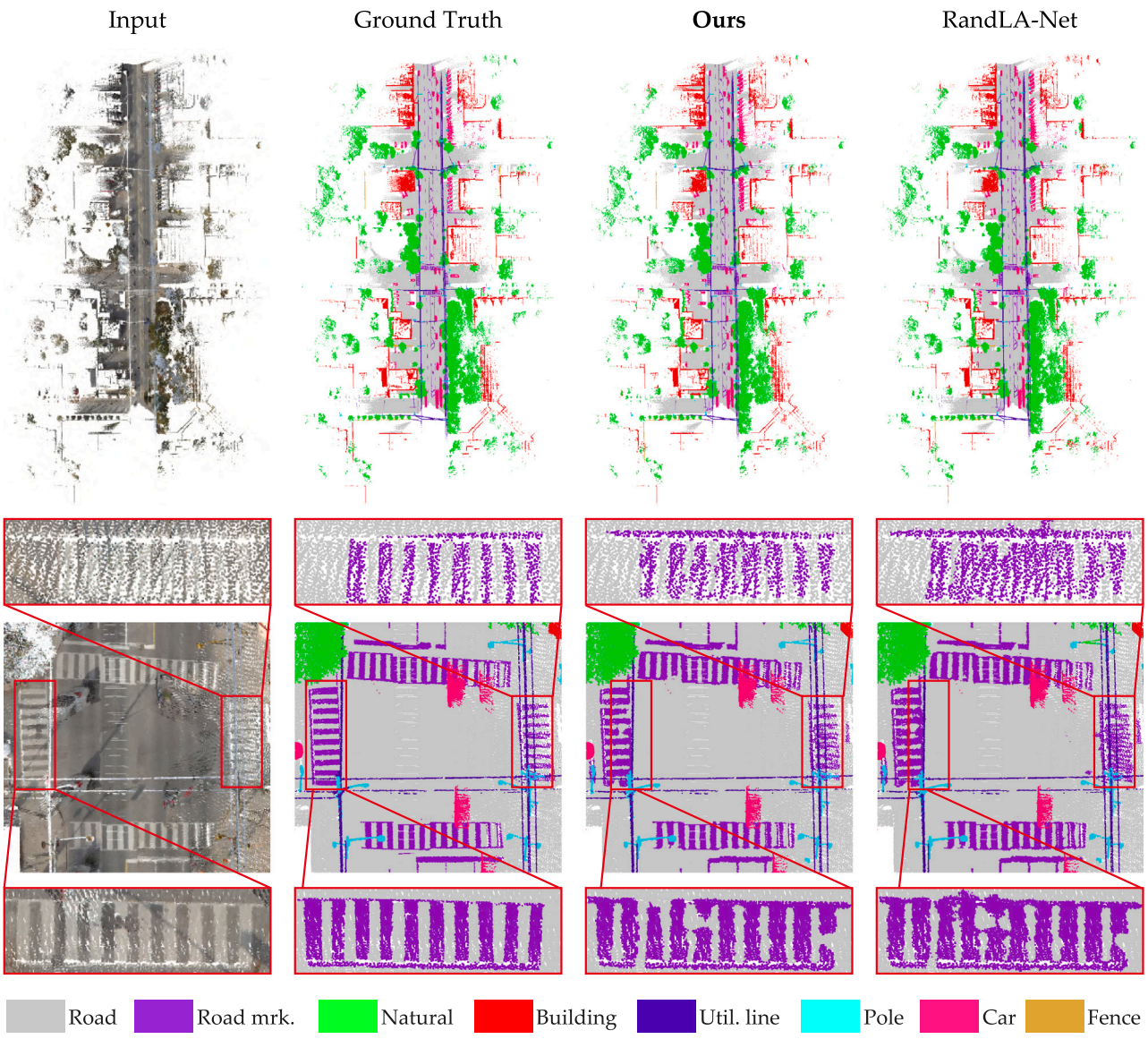


Fig. 8. Visual comparison of semantic segmentation results on Toronto3D dataset.

Table 6
Semantic segmentation results (IoU: %) of different baseline models with architectures on S3DIS (Area 5). MDS: Multi-level deep supervision.

Architecture	RandLA-Net	BAAF-Net	LACV-Net	PointNet++
U-Net	62.5	63.4	64.0	47.0
U-Net+	63.4 +0.9	62.9 -0.5	64.4 +0.4	49.1 +2.1
U-Net++	63.9 +1.4	64.0 +0.6	65.1 +1.1	50.8 +3.8
U-Net+^{d}	65.9 +3.4	65.6 +2.2	67.2 +3.2	54.9 +7.9
U-Next w/o MDS	66.7 +4.2	67.1 +3.7	68.0 +4.0	55.8 +8.8
U-Next (Ours)	68.3 +5.8	68.6 +5.2	69.9 +5.9	57.4 +10.4

long skip connection used in the U-Next architecture (Section 4.3.4). Finally, we examine the various applicability scenarios of different levels of U-Next in terms of performance and computational resource (Section 4.3.5). All subsequent models used in this study are trained on Areas 1–4 and 6, and tested on *Area 5* of the S3DIS (Armeni et al., 2016) dataset.

4.3.1. U-Next with different baselines

To verify the generalization ability of the proposed U-Next for different baseline models, we incorporate four representative approaches: PointNet++ (Qi et al., 2018), RandLA-Net (Hu et al., 2020), BAAF-Net (Qiu et al., 2021) and LACV-Net (Zeng et al., 2024b), into our U-Next architecture. For fair comparisons, we only use their local feature extraction modules as coding blocks in Eq. (1) and Eq. (2), despite several baselines introducing additional loss functions and architectural improvements. In Table 6, the quantitative results obtained by various baseline models are presented. It is evident that the semantic segmentation performance can be consistently improved (with an average improvement of 6.8% in mIoU scores) after incorporating the baseline model into our U-Next architecture. We also noticed that the improvement for PointNet++ is significant (10.4%), indicating that our U-Next architecture can significantly improve the feature learning capability of relatively weak backbones, by driving the network to iteratively fuse the multi-scale local feature maps generated by high-to-low and low-to-high sub-networks.

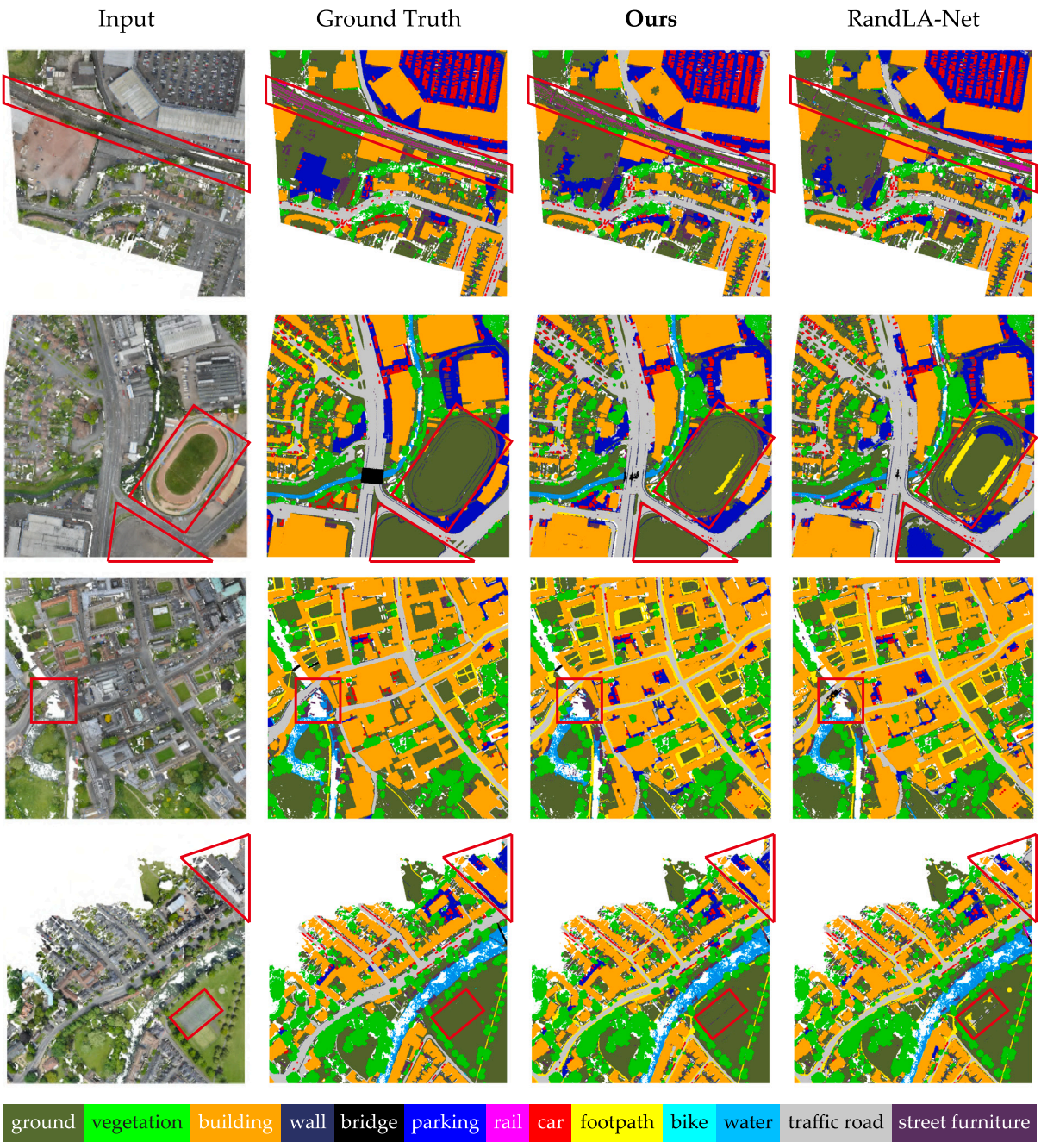


Fig. 9. Visual comparison of semantic segmentation results on SensatUrban dataset.

4.3.2. Varying U-Net-like architectures

On the other hand, to further demonstrate the effectiveness of the proposed U-Next compared to existing architectures, we take U-Net (Ronneberger et al., 2015), U-Net+, U-Net++ (Zhou et al., 2019) and the proposed U-Net+^{d} and U-Next as the framework for semantic segmentation of point clouds, respectively. As shown in the results in Table 6, all baseline methods achieve better performance when using the U-Next architecture, compared with that of other architectures. We also noticed that U-Net++, which is very successful in 2D medical images, failed to significantly improve the semantic segmentation performance of point cloud networks. By contrast, by stacking more U-Net L^1 subnetworks and leveraging multi-level deep supervision, the segmentation performance can be steadily improved.

Here, we additionally provide the qualitative comparison results achieved by RandLA-Net equipped with different architectures on S3DIS Area 5. As shown in Fig. 10, it can be seen that the proposed U-Next can segment the boundary areas smoothly and accurately. Meanwhile, several objects such as columns can be clearly segmented by the proposed U-Next, but not by U-Net or U-Net++, primarily because U-Net and U-Net++ only fuse horizontally with low-to-high feature maps, and cannot effectively explore sufficient multi-scale information for comprehensive representations.

4.3.3. Multi-level deep supervision

We proceed to validate the effectiveness of the proposed multi-level deep supervision. From the results in Table 6, U-Next with multi-level deep supervision further improves the mIoU score by 1.7% on average over U-Next without deep supervision. This further shows that stacking U-Net L^1 sub-networks and multiple-level deep supervision schemes may complement each other and jointly improve performance.

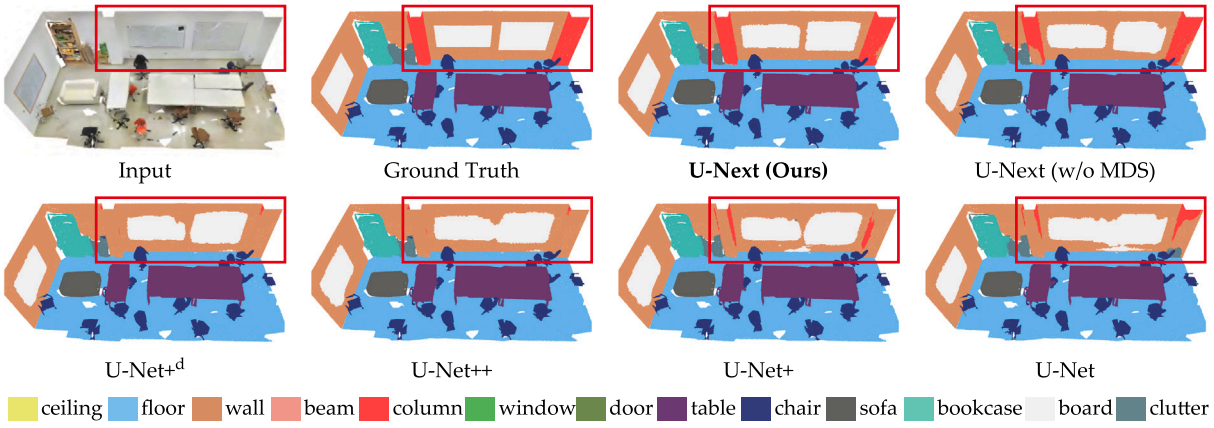
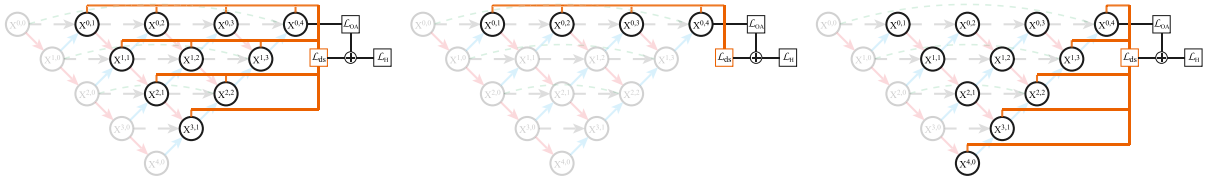


Fig. 10. Qualitative comparisons of RandLA-Net with different architectures on S3DIS (Area 5). MDS: Multi-level deep supervision.



(A) Multi-level deep supervision (B) Full-resolution deep supervision (C) Lateral deep supervision

Fig. 11. Illustration of the proposed multi-level deep supervision, the full-resolution deep supervision proposed by Zhou et al. (2019), and the lateral deep supervision proposed by Huang et al. (2020).

Table 7
Semantic segmentation results (%) of U-Next with different deep supervision on S3DIS (*Area 5*). DS.:deep supervision.

Architecture	OA (%)	mIoU (%)
w/o DS.	89.0	66.7
Full-resolution DS.	89.2	67.5
Lateral DS.	89.2	67.7
Multi-level DS.	89.5	68.3

Table 8
Semantic segmentation results (%) and parameters (millions) of U-Next with different skip connections on S3DIS (*Area 5*). conn.: connections.

Architecture	OA (%)	mIoU (%)	Parameters (millions)
w/o skip conn.	89.2	67.4	5.51
Dense skip conn.	89.4	67.9	5.66
Long skip conn.	89.5	68.3	5.61

Table 9
Semantic segmentation results (%) and parameters (millions) of different level of U-Next on S3DIS (*Area 5*).

Architecture	OA (%)	mIoU (%)	Parameters (millions)
U-Next L^4	89.5	68.3	5.61
U-Next L^3	89.1	66.1	1.43
U-Next L^2	86.8	60.6	0.31
U-Next L^1	81.0	49.6	0.05

In addition, the proposed U-Next introduces multi-level deep supervision from the perspective of optimization and learning in each decoder node. We note that U-Net++ (Zhou et al., 2019) only applies deep supervision on the generated full-resolution feature maps, and U-Net3+ (Huang et al., 2020) only applies deep supervision on the lateral nodes. To further verify the effectiveness of our multi-level deep supervision, we performed this ablation study (The comparison of the three is shown in Fig. 11). In this experiment, full-resolution deep supervision used in U-Net++ and lateral deep supervision used in U-Net3+ are integrated into U-Next. Table 7 shows the results of the quantitative analysis, from the results we can observe that deep supervision in full-resolution and lateral have superior performance compared to the network without using deep supervision, but has inferior performance compared to deep supervision in multi-level. This is mainly because the proposed multi-level deep supervision trains each U-Net L^1 sub-network independently, but the full-resolution deep or lateral supervision only trains on the original resolution scale or the rightmost nodes, the supervised sub-networks are limited.

4.3.4. Long skip connection

The proposed U-Next uses the design of long skip connections. It is noted that U-Net++ utilizes dense skip connections to connect all nodes at the same level. To verify the effectiveness of our design of long skip connections (and why not choose the design of U-Net++), we performed this ablation study. In this experiment, dense long skip connections are integrated into U-Next. As shown in Table 8, we can observe that although the network using dense skip connections have superior performance compared to networks without using long skip connections, it is comparable to the performance of using long skip connections. This may be because each sub-network can learn local features with different scales sufficiently, and the dense skip connections are not that necessary, but increase the risk of overfitting. In particular, dense skip connections introduce additional parameters and computational cost, resulting in redundant computations. Therefore, we finally adopt the vanilla long skip connections as used in the basic U-Net.

4.3.5. Different levels of U-Next

Here, we compare the performance and number of parameters of U-Next with different levels. Reducing the number of layers of U-Next can

Table 10
Model parameters (millions) and FLOPs (millions) of RandLA-Net with different architectures.

Architecture	Parameters	FLOPs
U-Net	4.99	31.35
U-Net+	5.24	31.96
U-Net++	5.39	33.02
U-Net+^{d}	5.51	33.79
U-Next	5.61	34.51

significantly reduce the inference time, but segmentation performance also degrades. As such, the level of U-Next should be a trade-off between computational costs and accuracy. As shown in Table 9, U-Next L^3 reduces the memory footprint (number of parameters) by 74.6%, while only reducing the mIoU score by 2.1%. More aggressive levels further reduce memory footprint, but at the cost of significant accuracy degradation. On the other hand, we also conducted similar experiments on U-Net with different levels, as shown in Table 9. It can be seen that the higher the level, the greater the performance improvement of our U-Next compared with the U-Net framework, primarily because more sub-networks are stacked as the level increase, which also means the more sufficient feature fusion.

4.4. Computational costs and model complexity

To investigate the complexity and computational costs of the proposed U-Next architecture, we report the number of parameters and floating point operations (FLOPs) for different architectures with RandLA-Net in Table 10. Not surprisingly, the proposed U-Next indeed has more trainable parameters and FLOPS, but not significant, since the addition of intermediate nodes does not bring massive trainable parameters, while most of the trainable parameters are generated in the intersection layer of the encoder–decoder architecture (due to the highest dimension in intermediate layers). Considering the segmentation performance and computational cost, the proposed U-Next architecture is meaningful and useful in practice.

Here we further provide detailed parameters in different layers of the architecture, as shown in Fig. 12. It can be seen that the deepest U-Net L^1 sub-network occupies the largest number of parameters compared to the other parts (about 75.76% of the total parameters), and the parameters of additional nodes in U-Next compared to the "U-Net" backbone only occupy a small part of parameters (about 10.57%). Therefore, the proposed U-Next does not introduce too many extra parameters compared to the original U-Net.

4.5. Compared with wide U-Net and U-Net++

Compared with the previous U-Net (Ronneberger et al., 2015) and U-Net++ (Zhou et al., 2019) framework, the proposed U-Next indeed introduces additional parameters. To verify that the performance improvement of U-Next is not attributed to additional parameters, but to the inherent advantages of the network architecture, we conducted the following experiments. We artificially increase the number of parameters of U-Net and U-Net++ by increasing the feature output dimension of the encoder and decoder. Specifically, the channel dimension of the features increases as: $(16+2 \rightarrow 64+4 \rightarrow 128+8 \rightarrow 256+16 \rightarrow 512+32)$. Table 11 shows the results of the quantitative analysis, it is clear that although introducing additional parameters can improve the segmentation performance, the segmentation performance of the proposed U-Next is still significantly superior to wide U-Net and U-Net++, which verifies that the performance improvements of the proposed U-Next are mainly due to the advantages of the network architecture.

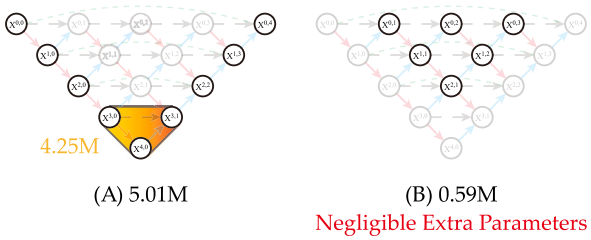


Fig. 12. Illustration of parameters (millions) in U-Next architecture. (A) shows parameters of "U-Net" backbone in U-Next. (B) shows parameters of additional nodes compared to "U-Net" backbone in U-Next.

Table 11
Semantic segmentation results (%) and parameters (millions) of different architectures on S3DIS (*Area 5*).

Architecture	OA (%)	mIoU (%)	Parameters (millions)
wide U-Net	87.4	63.2	5.64
wide U-Net++	87.8	64.9	6.09
U-Next	89.5	68.3	5.61

5. Conclusions

In this paper, we propose a U-Net-like framework for point cloud semantic segmentation, termed U-Next. The key to our U-Next framework is to stack a maximum number of U-Net L^1 sub-networks, ensuring minimal semantic gaps and better multi-scale feature representations. We also introduce a multi-level deep supervision mechanism to facilitate smooth gradient propagation and provide additional guidance at different levels of abstraction. Extensive experimental results on several large-scale benchmarks demonstrate the effectiveness of the proposed framework. In addition, our framework can be seamlessly integrated into existing point-based segmentation networks, resulting in improved segmentation accuracy. In future work, we plan to explore the applicability of U-Next to different modality data and tasks to further validate its generalizability and potential in diverse applications.

CRediT authorship contribution statement

Ziyin Zeng: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Qingyong Hu:** Writing – review & editing, Writing – original draft, Supervision, Methodology. **Zhong Xie:** Validation, Supervision. **Bijun Li:** Visualization, Validation, Supervision. **Jian Zhou:** Validation, Project administration. **Yongyang Xu:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Formal analysis.

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 work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2600300, Grant 2021YFB2600303, and Grant 2021YFB2600302.

Data availability

The code for our key contributions has been open sourced in https://github.com/zeng-ziyin/U-Next.

References

Ao, S., Guo, Y., Hu, Q., Yang, B., Markham, A., Chen, Z., 2022. You only train once: Learning general and distinctive 3D local descriptors. IEEE Trans. Pattern Anal. Mach. Intell..

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3D semantic parsing of large-scale indoor spaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2481–2495.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: IEEE/CVF International Conference on Computer Vision.

Blanc, T., El Beheiry, M., Caporal, C., Masson, J.-B., Hajj, B., 2020. Genuage: visualize and analyze multidimensional single-molecule point cloud data in virtual reality. Nature Methods 17 (11), 1100–1102.

Boulch, A., Saux, B.L., Audebert, N., 2017. Unstructured point cloud semantic labeling using deep segmentation networks. In: Eurographics Workshop on 3D Object Retrieval (3DOR).

Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N., 2016. A unified multi-scale deep convolutional neural network for fast object detection. In: European Conference on Computer Vision.

Chen, M., Hu, Q., Hugues, T., Feng, A., Hou, Y., McCullough, K., Soibelman, L., 2022. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset. In: British Machine Vision Conference.

Chen, C., Liu, D., Xu, C., Truong, T.-K., 2023. SAKS: Sampling adaptive kernels from subspace for point cloud graph convolution. IEEE Trans. Circuits Syst. Video Technol..

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306.

Chen, X., Ma, H., Wan, J., Li, B., Xia, T., 2017a. Multi-view 3D object detection network for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017b. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 834–848.

Chen, Y., Shen, C., Wei, X.-S., Liu, L., Yang, J., 2017c. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: IEEE/CVF International Conference on Computer Vision. ICCV.

Choy, C., Gwak, J., Savarese, S., 2019. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.-A., 2016. 3D deeply supervised network for automatic liver segmentation from CT volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Du, J., Cai, G., Wang, Z., Huang, S., Su, J., Marcato Junior, J., Smit, J., Li, J., 2021. ResDLPS-Net: Joint residual-dense optimization for large-scale point cloud semantic segmentation. ISPRS J. Photogramm. Remote Sens..

Du, Z., Ye, H., Cao, F., 2022. A novel local-global graph convolutional method for point cloud semantic segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Fan, S., Dong, Q., Zhu, F., Lv, Y., Ye, P., Wang, F.-Y., 2021. SCF-net: Learning spatial contextual features for large-scale point cloud segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018a. 3D semantic segmentation with submanifold sparse convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018b. 3D semantic segmentation with submanifold sparse convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Guan, S., Khan, A.A., Sikdar, S., Chitnis, P.V., 2019. Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. IEEE J. Biomed. Health Inf. 568–576.

Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R.R., Hu, S.-M., 2021. PCT: Point cloud transformer. Comput. Vis. Media.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3D point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (12), 4338–4364.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3d.net: A new large-scale point cloud classification benchmark. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Hengshuang, Z., Li, J., Jiaya, J., H.S., T.P., Vladlen, K., 2021. Point transformer. In: IEEE/CVF International Conference on Computer Vision.

Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A., 2022a. Sqn: Weakly-supervised semantic segmentation of large-scale 3D point clouds. In: European Conference on Computer Vision. Springer, pp. 600–619.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2021a. Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2022b. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. Int. J. Comput. Vis. 130 (2), 316–343.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2021b. Learning semantic segmentation of large-scale point clouds with random sampling. IEEE Trans. Pattern Anal. Mach. Intell..

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Huang, Q., Wang, W., Neumann, U., 2018. Recurrent slice networks for 3D segmentation of point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C., 2020. Virtual multi-view fusion for 3D semantic segmentation. In: European Conference on Computer Vision.

Landrieu, L., Raguet, H., Vallet, B., Mallet, C., Weinmann, M., 2017. A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds. ISPRS J. Photogramm. Remote Sens..

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. PointPillars: Fast encoders for object detection from point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Le, T., Duan, Y., 2018. PointGrid: A deep network for 3D shape understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. In: Artificial Intelligence and Statistics.

Li, H., Guan, H., Ma, L., Lei, X., Yu, Y., Wang, H., Delavar, M.R., Li, J., 2023a. MVPNet: A multi-scale voxel-point adaptive fusion network for point cloud semantic segmentation in urban scenes. Int. J. Appl. Earth Obs. Geoinf. 122.

Li, Y., Li, X., Zhang, Z., Shuang, F., Lin, Q., Jiang, J., 2022. DenseKPNET: Dense kernel point convolutional neural networks for point cloud semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2022.3162582.

Li, Y., Ma, L., Zhong, Z., Cao, D., Li, J., 2020. TGNet: Geometric graph CNN on 3-D point cloud segmentation. IEEE Trans. Geosci. Remote Sens..

Li, G., Muller, M., Thabet, A., Ghanem, B., 2019. DeepGCNs: Can GCNs go as deep as CNNs? In: IEEE/CVF International Conference on Computer Vision.

Li, X., Zhang, Z., Li, Y., Huang, M., Zhang, J., 2023b. SFL-NET: Slight filter learning network for point cloud semantic segmentation. IEEE Trans. Geosci. Remote Sens..

Liang, J., Du, Z., Liang, J., Yao, K., Cao, F., 2023. Long and short-range dependency graph structure learning framework on point cloud. IEEE Trans. Pattern Anal. Mach. Intell..

Liu, T., Ma, T., Du, P., Li, D., 2024. Semantic segmentation of large-scale point cloud scenes via dual neighborhood feature and global spatial-aware. Int. J. Appl. Earth Obs. Geoinf. 129, 103862.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019. Point-voxel CNN for efficient 3D deep learning. In: Advances in Neural Information Processing Systems (NeurIPS).

Ma, L., Li, Y., Li, J., Tan, W., Yu, Y., Chapman, M.A., 2021. Multi-scale point-wise convolutional neural networks for 3D object segmentation from LiDAR point clouds in large-scale environments. IEEE Trans. Intell. Transp. Syst..

Maturana, D., Scherer, S., 2015. VoxNet: A 3D convolutional neural network for real-time object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. ECCV.

Nie, D., Lan, R., Wang, L., Ren, X., 2022. Pyramid architecture for multi-scale processing in point cloud segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2018. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems (NeurIPS).

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H.A.A.K., Elhoseiny, M., Ghanem, B., 2022. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. In: Advances in Neural Information Processing Systems (NeurIPS).

Qiu, S., Anwar, S., Barnes, N., 2021. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Rong, M., Shen, S., 2023. 3D semantic segmentation of aerial photogrammetry models based on orthographic projection. IEEE Trans. Circuits Syst. Video Technol..

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Rusu, R.B., Blodow, N., Beetz, M., 2009. Fast point feature histograms (FPFH) for 3D registration. In: IEEE International Conference on Robotics and Automation.

Rusu, R.B., Cousins, S., 2011. 3D is here: Point cloud library (pcl). In: IEEE International Conference on Robotics and Automation. IEEE.

Samy, M., Amer, K., Eissa, K., Shaker, M., ElHelw, M., 2018. Nu-net: Deep residual wide field of view convolutional neural network for semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Saxena, S., Verbeek, J., 2016. Convolutional neural fabrics. In: Advances in Neural Information Processing Systems (NeurIPS).

Shuai, H., Xu, X., Liu, Q., 2021. Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation. IEEE Trans. Image Process..

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S., 2020. Searching efficient 3D architectures with sparse point-voxel convolution. In: European Conference on Computer Vision.

Tang, L., Zhan, Y., Chen, Z., Yu, B., Tao, D., 2022. Contrastive boundary learning for point cloud segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.-Y., 2018. Tangent convolutions for dense prediction in 3D. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. Segcloud: Semantic segmentation of 3D point clouds. In: International Conference on 3D Vision.

Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L., 2019. KPConv: Flexible and deformable convolution for point clouds. In: IEEE/CVF International Conference on Computer Vision.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks. In: IEEE/CVF International Conference on Computer Vision. ICCV.

Wang, H., Cao, P., Wang, J., Zaiane, O.R., 2022a. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J., 2019a. Graph attention convolution for point cloud semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, C., Ning, X., Sun, L., Zhang, L., Li, W., Bai, X., 2022b. Learning discriminative features by covering local geometric space for point cloud analysis. IEEE Trans. Geosci. Remote Sens. 60, 1–15. http://dx.doi.org/10.1109/TGRS.2022.3170493.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019b. Dynamic graph CNN for learning on point clouds. ACM Trans. Graph..

Wang, J., Wei, Z., Zhang, T., Zeng, W., 2016. Deeply-fused nets. arXiv:1605.07716.

Woo, S., Lee, D., Hwang, S., Kim, W.J., Lee, S., 2023. MKConv: Multidimensional feature representation for point cloud analysis. Pattern Recognit..

Xiao, X., Lian, S., Luo, Z., Li, S., 2018. Weighted res-unet for high-quality retina vessel segmentation. In: International Conference on Information Technology in Medicine and Education.

Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S., 2020. PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Ye, X., Li, J., Huang, H., Du, L., Zhang, X., 2018. 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In: European Conference on Computer Vision.

Yoo, S., Jeong, Y., Jameela, M., Sohn, G., 2023. Human vision based 3D point cloud semantic segmentation of large-scale outdoor scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Yu, T., Meng, J., Yuan, J., 2018. Multi-view harmonized bilinear network for 3D object recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Zeng, Z., Qiu, H., Zhou, J., Dong, Z., Xiao, J., Li, B., 2024a. PointNAT: Large-scale point cloud semantic segmentation via neighbor aggregation with transformer. IEEE Trans. Geosci. Remote Sens. 62, 1–18.

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J., Wu, W., 2022a. LEARD-Net: Semantic segmentation for large-scale point cloud scene. Int. J. Appl. Earth Obs. Geoinf..

Zeng, Z., Xu, Y., Xie, Z., Tang, W., Wan, J., Wu, W., 2024b. Large-scale point cloud semantic segmentation via local perception and global descriptor vector. Expert Syst. Appl..

Zeng, Z., Xu, Y., Xie, Z., Wan, J., Wu, W., Dai, W., 2022b. RG-GCN: A random graph based on graph convolution network for point cloud semantic segmentation. Remote Sens..

Zhan, L., Li, W., Min, W., 2023. FA-ResNet: Feature affine residual network for large-scale point cloud segmentation. Int. J. Appl. Earth Obs. Geoinf..

Zhang, J., Jiang, Z., Qiu, Q., Liu, Z., 2024. TCFAP-Net: Transformer-based cross-feature fusion and adaptive perception network for large-scale point cloud semantic segmentation. Pattern Recognit. 154, 110630.

Zhao, H., Jiang, L., Fu, C.-W., Jia, J., 2019. PointWeb: Enhancing local neighborhood features for point cloud processing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Zhou, H., Feng, Y., Fang, M., Wei, M., Qin, J., Lu, T., 2023. AGConv: Adaptive graph convolution on 3D point clouds. IEEE Trans. Pattern Anal. Mach. Intell..

Zhou, C., Shu, Z., Shi, L., Ling, Q., 2024. Semantic segmentation for large-scale point clouds based on hybrid attention and dynamic fusion. Pattern Recognit. 156, 110798.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 1856–1867.

Zhu, Q., Du, B., Turkbey, B., Choyke, P.L., Yan, P., 2017. Deeply-supervised CNN for prostate segmentation. In: International Joint Conference on Neural Networks.