Fusion-SUNet: Spatial Layout Consistency for 3D Semantic Segmentation

Maryam Jameela, Gunho Sohn, Sunghwan Yoo
Department of Earth and Space Science and Engineering,York University
Toronto, ON M3J 1P3, Canada
{maryumja,gshon,jacobyoo}@yorku.ca

Abstract

The paper discusses the need for a reliable and efficient computer vision system to inspect utility networks with minimal human involvement, due to the aging infrastructure of these networks. We propose a deep learning technique, Fusion-Semantic Utility Network (Fusion-SUNet), to classify the dense and irregular point clouds obtained from the airborne laser terrain mapping (ALTM) system used for data collection. The proposed network combines two networks to achieve voxel-based semantic segmentation of the point clouds at multi-resolution with object categories in three dimensions and predict two-dimensional regional labels distinguishing corridor regions from non-corridors. The network imposes spatial layout consistency on the features of the voxel-based 3D network using regional segmentation features. The authors demonstrate the effectiveness of the proposed technique by testing it on $67km^2$ of utility corridor data with average density of $5pp/m^2$, achieving significantly better performance compared to the state-of-the-art baseline network, with an F1 score of 93% for pylon class, 99% for ground class, 99% for vegetation class, and 98% for powerline class.

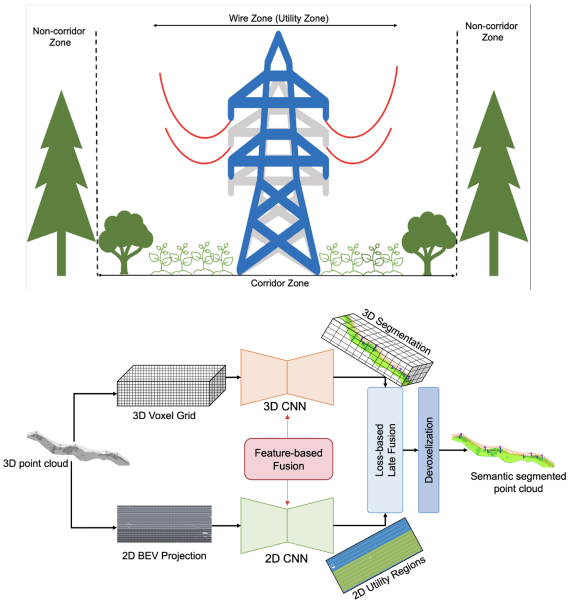


Figure 1. SUNet is a multi-dimensional and multi-resolution network that imposes the spatial layout consistency (1a) through a 2D bird's-eye view (BEV) of utility regions on the outcomes of 3D segmentation network via loss-based late fusion (1b).

1. Introduction

The paper focuses on the development of a computer vision system that can inspect utility networks with minimal human involvement. The aging infrastructure of these networks has made it essential to have an efficient system for inspecting and managing them [17]. Although unmanned aerial vehicles (UAVs) have been used for utility inspections, there are still challenges in using them to collect data across entire utility corridors. As a result, Airborne Laser Terrain Mapping (ALTM) has become the primary data collection platform [41] [28]. However, labeling semantic features in point clouds using visual perception tasks is still a challenging and expensive task that often requires manual labor and is prone to errors. As a result, there is a significant need to automate post-data acquisition procedures to reduce the level of user involvement and improve efficiency [16] [37]. Deep Convolutional Neural Networks (DCNNs) have recently shown significant improvements in computer vision, particularly for semantic segmentation using point clouds [35] [29] [30]. However, existing networks have not fully exploited the spatial arrangement of infrastructure, especially for utility corridors, nor have they embedded spatial layout consistency for global context. This limitation motivated the authors to propose a network with hierarchical spatial regularity that can be generalized for standard layout segmentation problems. The proposed network, called Fusion-Semantic Utility Network (Fusion-SUNet), is a multi-dimensional and multi-resolution network that combines two networks to classify point clouds at multi-resolution with object categories in three dimensions and

predict two-dimensional regional labels distinguishing corridor regions from non-corridors, as shown in Figure 1.
The network imposes spatial layout consistency on the features of voxel-based 3D network through a multi-dimensional feature fusion module using regional segmentation features. The authors tested the performance of the Fusion-SUNet using utility corridor data and demonstrated that it significantly improves performance compared to the state-of-the-art baseline network, achieving high F1 scores for various classes. The following sections of the paper discuss related work, the proposed methodology, experiments, and results.

2. Literature Review

This section aims to explore the spatial regularities and layout consistency present in various semantic scenes. These regularities can provide important information for identifying and understanding objects of interest in a scene. However, these regularities have not been fully utilized for visual perception tasks. In this research, we investigate the potential of leveraging spatial layout consistency in utility corridors to perform semantic segmentation of the utility network.

2.1. Spatial Layout

The concept of spatial layout and the relationship between objects in a scene has been explored in cognitive science, architectural design, and civil engineering [1]. The relationship between objects provides context and aids in classification. Convolution neural networks have been designed to extract these spatial relationships, but embedding spatial regularities is necessary to learn global context. The railway lane extractions, road lane detection, and 3D building modeling are prime examples of utilizing spatial regularities to improve results [6, 13]. Previous studies have demonstrated the importance of spatial relationships in detecting small objects that may be overlooked [32]. This motivates our study to use the spatial layout and object relationships in utility corridor regions to highlight the importance of embedding spatial layout consistency in our network.

2.2. Utility Network Layout

The global guidelines for electric hydro companies require the establishment of utility transmission zones that ensure the safety of infrastructure, prevent vegetation encroachment, and consider residential areas [23]. These transmission zones' size and shape are dictated by voltage and specific regulations of each country, with three primary zones being the utility zone, corridor zone, and non-corridor zone. The utility zone can only have low vegetation, while the corridor zone has trees up to 3-5 meters high, and the non-corridor zone can have tall trees [3] [10]. In this paper, we combined the utility and corridor zones to create a single corridor zone to explore the spatial layout consistency of these zones and address the semantic segmentation problem for utility networks. We reviewed existing literature on utilizing hierarchical relationships for visual perception, such as detecting human motion, to understand the limitations and benefits of various semantic segmentation networks and establish a baseline for their deep neural network [36] [25].

2.3. Semantic Segmentation

In recent years, significant advancements have been made in the area of semantic segmentation for 3D point clouds, mainly due to the widespread use of deep learning in computer vision and artificial intelligence. A lot of research has focused on utilizing intrinsic, extrinsic, and deep features to classify each point in the point cloud with an enclosing object that is relevant [9]. In prior sections, we have identified that enforcing spatial layout consistency based on real-world context is a significant challenge in semantic segmentation networks, as well as global context embedding. Despite this, we will discuss current semantic segmentation approaches which fall into three categories: statistical segmentation, classification networks based on machine learning, and segmentation networks based on deep learning.

2.3.1 Geometrical Segmentation

Traditionally, segmentation of utility corridors has been viewed as a purely geometric problem. As a result, algorithms were designed to extract lines, group point clouds, and categorize them using features such as neighborhood votes, density, and elevation-based attributes [20]. However, these methods have several limitations, including the need for extensive preprocessing, feature engineering, and domain expertise. Additionally, these systems require multiple filtering steps to segment the objects [14]. Furthermore, these techniques are not robust for raw, large-scale, dense 3D point clouds [16].

2.3.2 Machine Learning based Utility Classification

Previous studies have successfully classified utility objects, reconstructed transmission lines, and extracted features with various machine learning algorithms like random forest [17], support vector machines [37], decision trees [19], and balanced/unbalanced learning [15]. These models mainly rely on handcrafted features based on geometric characteristics and either 3D voxels or 2D projections to improve classification accuracy. However, they face limitations when applied to 3D large-scale datasets.
In contrast, deep learning has the potential to automatically learn features and interpret data for any computer vision task, making it a promising approach for utility network management. Deep learning has enabled researchers

to work across different domains without extensive domain knowledge, and it has allowed the development of generalizable solutions that can be applied across various datasets and sensors.

2.3.3 Deep Learning-based Segmentation Networks

Various input representations have been used to train deep neural networks for semantic segmentation of point clouds, including 3D voxels and 2D multiview projections. While these representations offer an effective performance boost, they also introduce quantization errors. In the last decade, a new batch of methods that use raw point clouds as input have been introduced, starting with PointNet [29]. These methods, including PointNet++ [30], KPConv with continuous kernels [35], and the state-of-the-art network RandLA [9], have demonstrated comparable performance on most segmentation benchmarks, such as Semantic3D [5], SensatUrban [8], and DublinCity [42]. However, as previously discussed, none of these methods have taken advantage of the spatial regularity found in utility infrastructure. This limitation inspired us to propose Fusion-SUNet, which is an extension of SUNet [12] that utilizes regional results to impose spatial layout consistency. The Fusion-SUNet approach fuses features from the regional network at every layer of the decoder, providing spatial guidelines to improve performance on a deeper level.

3. Methodology

Fusion-SUNet is a multi-dimensional and multi-resolution network that incorporates spatial layout consistency between the layout and objects of a scene. It comprises two networks: a two-dimensional regional prediction network [31] that constrains the predictions of a three-dimensional network through multi-dimensional deep feature fusion based module and loss-based late fusion. The network architecture enables multi-dimensional fusion at different depths, allowing it to incorporate regional information from the 2D network to improve the 3D network's prediction accuracy [7] [38].

3.1. 3D Semantic Segmentation Pipeline

Figure 2 shows an overview of the network architecture of Fusion-SUNet. The 3D pipeline consists of a U-shaped multiresolution encoder-decoder network that learns features from a three-dimensional voxel grid to assign semantic labels for 3D object classes. The 3D network has four encoded feature maps E_1, E_2, E_3, and E_4, which are used to create four decoded feature maps D_1, D_2, D_3, and D_4 through serial operations that include additive attention applied on D_{n-1} and E_n. These output feature maps P^A_{3D} from additive attention modules are then fused using a feature-based fusion module with 2D features $P^A_{2D\_3D}$ from the regional segmentation layer. The fused features $P^{A"}_{3D}$ are passed through two 3x3x3 convolutions, followed by batch normalization and rectified linear activation, resulting in four decoded feature maps of dimensions $\frac{H}{2l} \times \frac{W}{2l} \times \frac{D}{2l} \times 32l$ for output D at level l. These feature maps represent the deep multi-resolution output that provides a multi-receptive field for segmenting objects and areas of different sizes. The feature maps are then passed through a multi-resolution feature aggregation module to incorporate knowledge and context from all levels. It delivers a confidence score against each class. This is then passed through a loss-based late fusion module to refine and constrain the predictions using spatial layout consistency and back-propagate the loss to better learn deep features.

3.1.1 Additive Attention Module

The module creates multiple attention coefficients for each class, which helps to filter out irrelevant information from the feature maps [24]. These attention gates take input from the previous layer of the decoder (D_1) and the encoder feature map (E_2) at the same level, and generate attention coefficients (α) to combine with the feature map (E_2) using element-wise summation.

3.1.2 Multi-resolution Feature Aggregation Module

This module which combines the feature maps from four different levels of the decoder (D_1, D_2, D_3, D_4) by upsampling and concatenating them. The resulting aggregated feature map is then passed through a 1x1x1 convolution. This module is useful for producing a feature map that corrects any errors introduced during the upsampling process in the decoder, especially for predicting minority classes.

3.2. 2D BEV Pipeline

Incorporating global context and a larger receptive field into a semantic segmentation network is a challenging task for making accurate spatial predictions. Humans rely heavily on understanding the global context when perceiving scene semantics. However, our 3D segmentation network only considers the local context and misses important details of the scene that are needed to encode the spatial arrangement and overall correlation of objects. To overcome this limitation, we employ a two-dimensional (2D) bird's-eye-view (BEV) pipeline that uses a feature-based fusion module and a loss-based late fusion to combine the missing global contextual information. The 2D network is a multi-resolution encoder-decoder shown in Figure 2 as a 2D BEV pipeline [31]. It consists of four feature maps E'_1, E'_2, E'_3, E'_4, which are used to construct four output maps D'_1, D'_2, D'_3, D'_4. This network takes in a Bird's Eye View (BEV) representation of a complete 3D point cloud scene

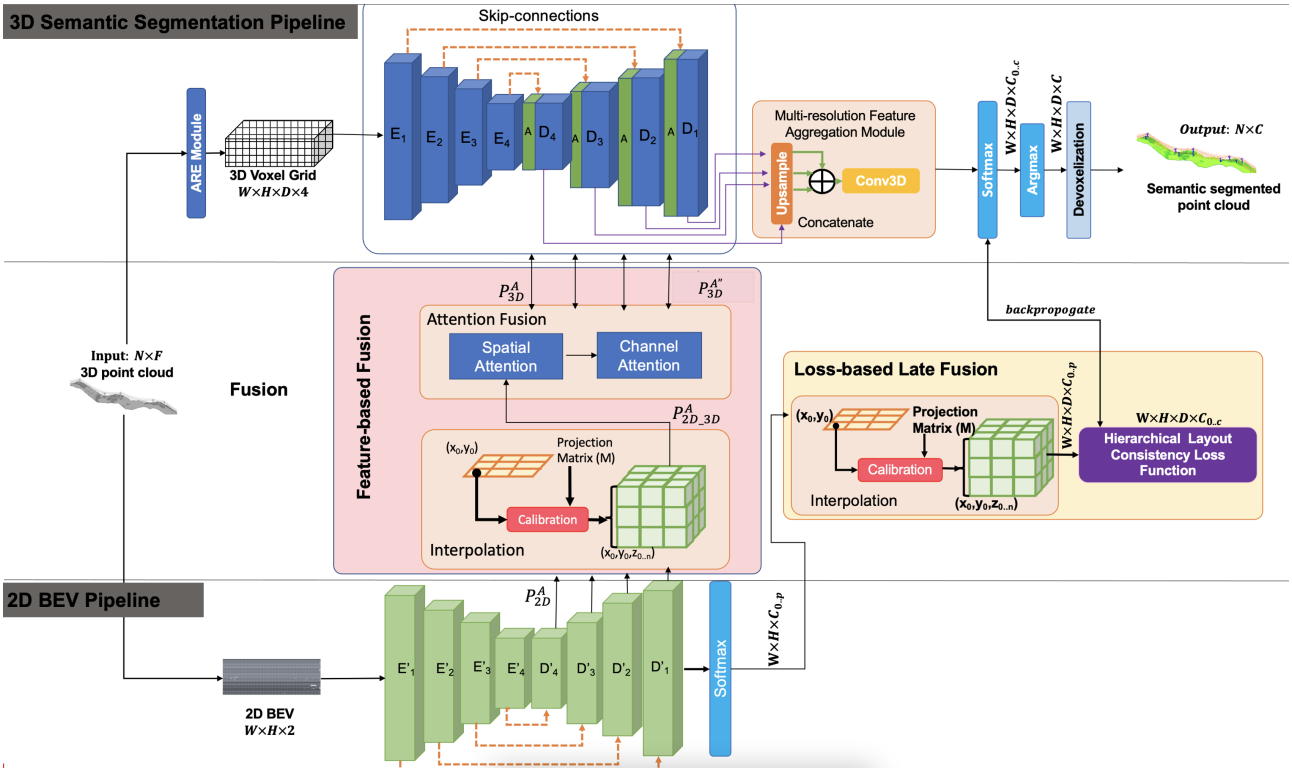


Figure 2. The Overall architecture of Fusion-SUNet: Point cloud is preprocessed into voxel grid and BEV. These are separately processed by a multi-resolution 3D semantic segmentation pipeline and 2D BEV pipeline for regional prediction to impose the spatial layout consistency on 3D objects through feature-based fusion and loss-based late fusion.

and predicts a regional classes probability tensor of shape $W \times H \times C_p$.

3.3. Fusion

Fusion pipeline have two key modules, Feature based middle fusion and loss based late fusion. SUNet have proved the importance of loss based late fusion already. We will discussed feature based fusion in this section.

3.3.1 Feature Interpolation Module

It is a simple module that takes the projection matrix between 3D voxels and 2D BEV and converts the 2D features map for regional network's layers of shape $\frac{H}{2l} \times \frac{W}{2l} \times 32l$ into a 3D representation of shape $\frac{H}{2l} \times \frac{W}{2l} \times \frac{D}{2l} \times 32l$ represented as P_{2D_3D} by exploring the one-to-many relationship between the two representations.

3.3.2 Attention Fusion Module

The module is divided into two attention modules: spatial attention and channel-based attention as shown in Figure 3. The spatial attention mechanism generates an attention-focused feature map by interpolating a 2D regional network to a 3D feature map and concatenating it with the 3D feature map from the additive attention map output to provide a feature map for a decoder level in the 3D network. The spatial attention mechanism imposes layout constraints on the features to facilitate the learning of objects of interest by taking advantage of spatial consistency at deeper levels.

- The attention fusion module takes input from the feature interpolation module $P_{2D\_3D_l}$ and the output of the additive attention module from the 3D segmentation pipeline $P^A_{3D_l}$.
- The attention fusion module performs spatial attention using both feature maps, using element-wise summation and ReLU activation as the final step, followed by element-wise multiplication with the additive attention map output.
- The channel attention map is generated from the interpolated 2D to 3D regional feature map using max-pooling and average pooling operations, followed by a shared MLP and element-wise summation to generate a 1D channel attention map of size 1 x 1 x 1 x 32l after a sigmoid operation.
- The output of the channel attention mechanism

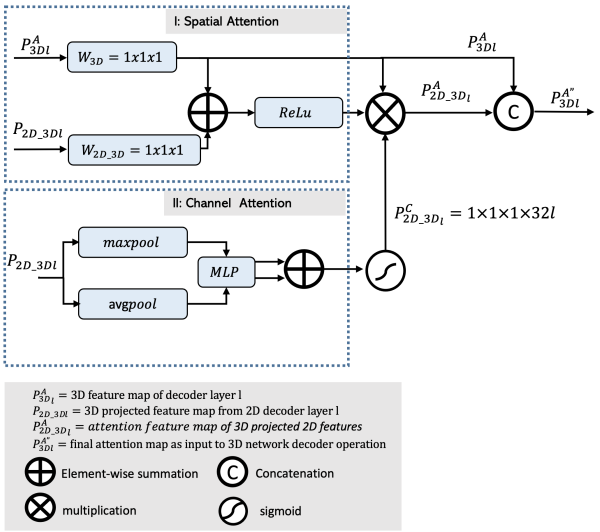


Figure 3. Attention fusion module to impose spatial layout consistency through fusion of features also known as middle fusion.

$P^C_{2D\_3D_l}$ is multiplied element-wise with the output of the spatial attention map $P^A_{2D\_3D_l}$ and later concatenated with the feature map $P^A_{3D_l}$ from the 3D network.

- The final output of the feature fusion module is denoted as $P^{A"}_{3D_l}$ and is sent to the 3D network for decoder operations involving two 3x3x3 convolution batch normalization, ReLU, and upsampling.

3.3.3 Loss based Late Fusion:

It uses the hierarchical layout consistency loss function as used in SUNet as shown in equation 1 to impose the spatial layout consistency on outcome. M is total number of samples, C total number 3D classes and P is total number of 2D classes.

$$L = \frac{1}{M}\sum_{p=1}^{P}\sum_{c=1}^{C}\sum_{m=1}^{M}(w_p w_c)\times(y_m^{P_c}\times(log(h_\Theta(x_m,c,p)))) \tag{1}$$

3.4. Voxelization and BEV Projection

The input representation for our segmentation network is a voxel grid, which is created by pre-processing the raw point cloud and calculating a mean value for all the points that fall within each 3D voxel. This voxel grid provides a balance between efficiency and effectiveness, depending on the selected voxel size. Additionally, the network maintains a projection matrix from the voxel grid to the raw point cloud to enable easy projection.
In contrast, a bird's eye view (BEV) is a 2D representation of a 3D point cloud. Our 2D BEV pipeline utilizes the XY-projection of a 3D scene, where each pixel represents the residing points. The XY-projection yields the most optimal BEV for extracting global context for regional and object prediction. Our projection matrix between the BEV and 3D voxel grid provides projection compatibility between feature spaces to combine the 2D and 3D predictions, resulting in better utilization of spatial layout consistency.

4. Experiments and Results

Our research involved a comparative analysis to demonstrate the significance of the multi-dimensional feature fusion module. We evaluated the performance of our proposed approach on a test set, which included four key classes that are highly relevant to the utility industry: ground, pylon, powerline, and vegetation. These classes play a critical role in predictive maintenance of utility networks.

4.1. Dataset

In our experimental setup, we used a Riegl Q560 laser scanner to capture data from an area of $67km^2$ in Steamboat Springs, Colorado, USA. The collected data was divided into two sets, training and testing as shown in figure 4. The first $8km^2$ of the dataset was used for testing while the remaining data was used to train the network. The dataset contained 67 non-overlapping scenes, with each scene containing over two million points and an average density of $5ppc/m^2$. We manually labeled the data using Terrasolid point cloud processing software to generate the ground truth labels. The training dataset consisted of five classes: powerline, pylon, low vegetation, ground, and medium-high vegetation. To simplify the ground class, we merged the low vegetation class with the ground class as most of the ground was covered by low vegetation. Moreover, we named our regional classes based on existing utility community literature [3, 10]. The corridor region comprised of powerlines and pylons located within a 3-10 meter range from the pylon, while the non-corridor region was located outside this range.

4.2. Experimental Configuration

To conduct our experiments, we pre-trained our 2D regional prediction network on half of the scenes for global regional prediction of spatial layout. Each scene was split into four subscenes based on GPS time of flight line and projected onto a 2D BEV grid of size 640×640 with a pixel size of $1m^2$. The network was trained with a batch size of 1 and 100 epochs using K-cross validation to prevent overfitting, and data augmentation was performed using horizontal flip, vertical flip, and random rotation. The input size for the 2D network was $640 \times 640 \times 2$, and the output was $640 \times 640 \times 3$ representing confidence scores for regional

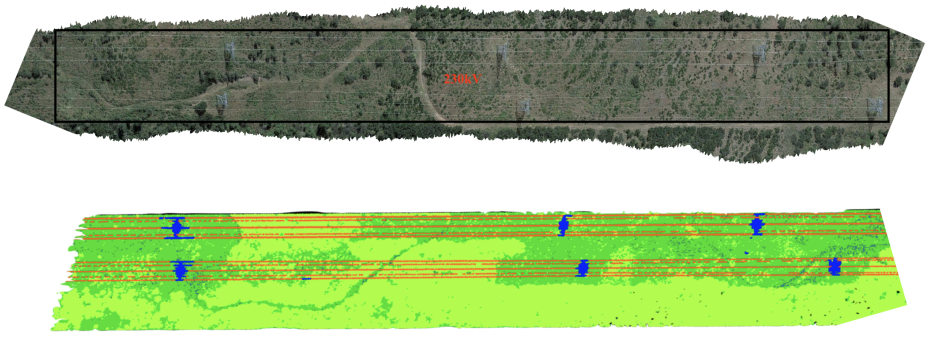


Figure 4. Visualization of Data from Google Maps and 3D point cloud acquired from Steamboat Springs, Coloardo, USA

classes. Training was performed on two GPU RTX 6000, taking 4-5 hours, with inference taking about 30 seconds.

For the voxel grid, we generated it over each subscene with a size of $640 \times 640 \times 448$ and a voxel size of $1m^3$. Each batch $32 \times 32 \times 448 \times 4$ contained the maximum elevation of the entire scene to provide the network with a complete view and better handle vertical context. Feature channels included absolute and relative elevation, the number of occupancy points, and the number of returns, selected based on a feature engineering study discussed in SUNet [12]. Fusion-SUNet fuses features on all four levels of decoder from the 2D regional network to the 3D voxel-based network. It outputs $32 \times 32 \times 448 \times 5$, representing confidence scores against 3D classes (background, pylon, powerline, vegetation, and ground). The final prediction assigns a true label based on the highest confidence score and projects voxel labels onto points using the projection matrix. Fusion-SUNet was trained on two GPU RTX 6000 for 100 epochs, taking 48-60 hours, with inference taking about 2-3 minutes.

4.3. Evaluation Matrices

The paper reports the evaluation metrics in terms of F1 score for each of the four classes: ground, pylon, powerline, and vegetation. The F1 score is calculated as the harmonic mean of precision and recall. The precision measures how many of the predicted positive classes are actually positive, while recall measures how many of the actual positive classes are correctly predicted as positive. The F1 score provides an overall measure of the model's performance, taking both precision and recall into account.

4.4. Results

We conducted experiments to compare Fusion-SUNet with various versions of SUNet [12], Attention 3D [24], and pre-trained RandLA [9]. The results showed that the feature fusion module in Fusion-SUNet provides significant advantages, as demonstrated by the higher recall and F1 score in the pylon class as shown in figure 5. This indicates that incorporating global scale spatial layout context on the feature level is beneficial. Our findings also suggest that point-based RandLA has difficulty generalizing over the veg and pylon classes due to a lack of spatial context. To improve RandLA's results, we could integrate spatial layout consistency through middle and loss-based fusion in the future. We chose a voxel-based network due to its comparable quality and performance in a shorter time frame, and our inference is 10 times faster than point-based networks such as RandLA.

We also concluded that our network outperforms existing commercial software products, such as Terrasolid [34] and Cobravision [33] , which require manual labeling of powerlines and pylons or need additional information of utility network maps to perform predictive analysis.

5. Conclusion

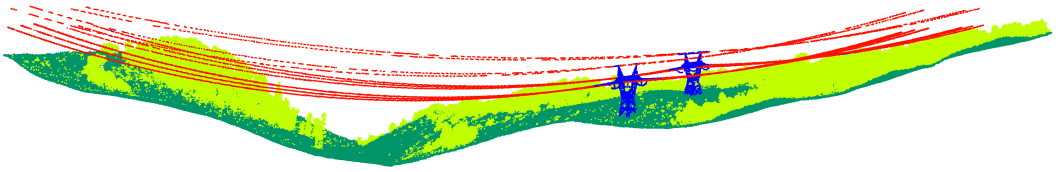
In our study, we have shown that the Fusion-SUNet model, which incorporates a middle fusion module for multi-dimensional feature fusion and a late fusion module based on loss, can effectively embed spatial layout. Our experiments have confirmed that the integration of hierarchical layout spatial consistency with a coarse-to-fine strategy can improve the performance of deep semantic segmentation models for predictive analysis. Moving forward, our research will aim to test the generalizability of this network across different sensors, and to conduct a detailed ablation study.

6. Acknowledgment

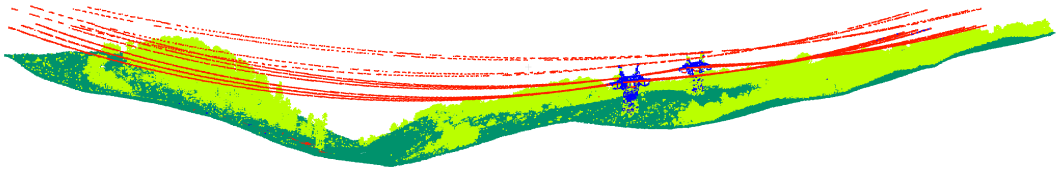
This research project has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC)'s Collaborative Research and Development Grant (CRD) – 3D Mobility Mapping Artificial Intelligence (3DMMAI) and Teledyne Geospatial Inc. We'd like to thank Alvin Poernomo (Machine Learning Developer), Hamdy Elsayed (Innovation Manager) and Chris Verheggen (SVP R& D).

Table 1. Comparative study of Fusion-SUNet to SUNet + MFA (multi feature aggregation) module, SUNet + MFA + FS (feature smoothing), Atention UNet and RandLA on recall (Rec), precision (Prec) and F1 score (F1).

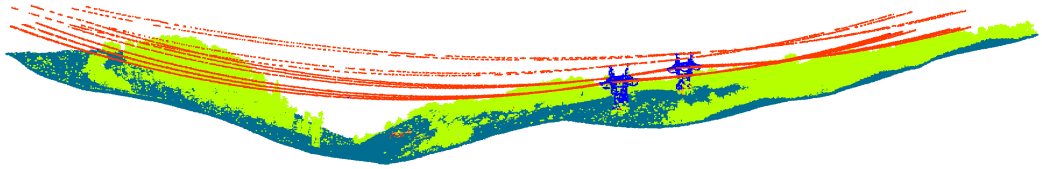
Methods	Pylon			Ground			Veg			Powerline		
	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
3D AUNet (baseline)	72.0	84.0	77.0	93.0	94.0	93.0	97.0	97.0	97.0	84.0	98.0	91.0
RandLA	75.0	92.0	83.5	95.0	96.0	95.5	76.0	87.0	81.5	86.5	98.0	91.3
SUNet+MFA+FS	78.0	97.0	87.0	**100.0**	99.0	**100.0**	99.0	99.0	99.0	**99.0**	97.0	98.0
SUNet+MFA	82.0	96.0	89.0	99.0	99.0	99.0	99.0	99.0	99.0	98.0	99.0	99.0
Fusion-SUNet	**94.0**	**92.0**	**93.0**	99.0	**99.0**	99.0	**99.0**	**99.0**	**99.0**	98.0	**99.0**	**99.0**



Groundtruth



SUNet + ARE Module + MFA



Fusion-SUNet

Figure 5. Visualization of results; Groundtruth, SUNet + ARE + Multiresolution feature aggregation module and Fusion-SUNet (ours). Blue: pylon, red: powerline, green: high vegetation and dark-green: ground

References

[1] Martin Brosamle and Christoph Holscher. Architects seeing through the eyes of building users, a qualitative analysis of design cases. *2007, International Conference on Spatial Information Theory (COSIT'07)*, pages 8–13, 01 2007. 2

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI,2016)*, 9901:424–432, 2016.

[3] Electric Power Research Institute (EPRI). *Vegetation Management - Best Practices for Transmission Corridors*. Electric Power Research Institute, 2012. 2, 5

[4] Haiyan Guan, Yongtao Yu, Jonathan Li, Zheng Ji, and Qi Zhang. Extraction of power-transmission lines from vehicle-borne lidar data. *International Journal of Remote Sensing*, 37(1):229–247, 2016.

[5] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan Dirk Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark. *CoRR*, abs/1704.03847, 2017. 3

[6] Mandar Haldekar, Ashwinkumar Ganesan, and Tim Oates.

Identifying spatial relations in images using convolutional neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3593–3600, 2017. 2

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. 3

[8] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. *CoRR*, abs/2009.03137, 2020. 3

[9] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. *CoRR*, abs/1911.11236, 2019. 2, 3, 6

[10] International Society of Arboriculture (ISA). *Guidelines for Vegetation Management Near Power Lines*. International Society of Arboriculture, 2017. 2, 5

[11] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge. *CoRR*, abs/1802.10508, 2018.

[12] Maryam Jameela and Gunho Sohn. Spatial layout consistency for 3d semantic segmentation, 2023. 3, 6

[13] Wang-Gyu Jeon and Eui-Myoung Kim. Automated reconstruction of railroad rail using helicopter-borne light detection and ranging in a train station. *Sensors and Materials*, 31:3289, 10 2019. 2

[14] Jaehoon Jung, Erzhuo Che, Michael J. Olsen, and Katherine C. Shafer. Automated and efficient powerline extraction from laser scanning data using a voxel-based subsampling with hierarchical approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:343–361, 2020. 2

[15] Yoonseok Jwa and Gunho Sohn. A multi-level span analysis for improving 3d power-line reconstruction performance using airborne laser scanning data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38:97–102, 09 2010. 2

[16] Yoonseok Jwa, Gunho Sohn, and Heungsik Kim. Automatic 3d powerline reconstruction using airborne lidar data. *IAPRS*, 38:105–110, 01 2009. 1, 2

[17] Heungsik Kim and Gunho Sohn. 3d classification of powerline scene from airborne laser scanning data using random forests. *IAPRS*, 38:126–132, 09 2010. 1, 2

[18] Heungsik Kim and Gunho Sohn. Random forests based multiple classifier system for power-line scene classification. volume XXXVIII-5/W12, 08 2011.

[19] Heungsik Kim and Gunho Sohn. Point-based classification of power line corridor scene using random forests. *Photogrammetric Engineering and Remote Sensing*, 79:821–833, 09 2013. 2

[20] Yuee Liu, Zhengrong Li, Ross Hayward, Rodney Walker, and Hang Jin. Classification of airborne lidar intensity data using statistical analysis and hough transform with application to power line corridors. In *2009 Digital Image Computing: Techniques and Applications*, pages 462–467, 2009. 2

[21] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam Gyu Cho, Seong Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 891–898. IEEE Computer Society, Sept. 2014. Publisher Copyright: © 2014 IEEE.; 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014 ; Conference date: 23-06-2014 Through 28-06-2014.

[22] Zhixiong Nan, Jizhi Peng, Jingjing Jiang, Hui Chen, Ben Yang, Jingmin Xin, and Nanning Zheng. A joint object detection and semantic segmentation model with cross-attention and inner-attention mechanisms. *Neurocomputing*, 463:212–225, 2021.

[23] National Grid Transco UK, Wales, and USA. Design guidelines for development near high-voltage overhead lines. https://www.nationalgrid.com/uk/electricity-transmission/publications/design-guidelines-development-near-high-voltage-overhead-lines, 11 12 2022. Accessed: DD Month YYYY. 2

[24] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. 3, 6

[25] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Computer Vision and Pattern Recognition*, pages 459–468. IEEE, 2018. 2

[26] Jizhi Peng, Zhixiong Nan, Linhai Xu, Jingmin Xin, and Nanning Zheng. A deep model for joint object detection and semantic segmentation in traffic scenes. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[27] Kirsten Petras, Sanne ten Oever, Christianne Jacobs, and Valerie Goffaux. Coarse-to-fine information integration in human vision. *NeuroImage*, 186:103–112, 2019.

[28] S. Pu, L. Xie, M. Ji, Yongliang Zhao, W. Liu, L. Wang, F. Yang, and D. Qiu. Real-time powerline corridor inspection by edge computing of uav lidar data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W13:547–551, 06 2019. 1

[29] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 1, 3

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 1, 3

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 3

[32] Benjamin Rosman and Subramanian Ramamoorthy. Learning spatial relationships between objects. *I. J. Robotic Res.*, 30:1328–1342, 10 2011. 2

[33] Cobravision Team. Cobravision. https://www.cobravision.com, 2023. Accessed on March 5, 2023. 6

[34] Terrasolid Team. Terrasolid. https://www.terrasolid.com, 2023. Accessed on March 5, 2023. 6

[35] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *CoRR*, abs/1904.08889, 2019. 1, 3

[36] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition*, pages 1653–1660. IEEE, 2014. 2

[37] Yanjun Wang, Qi Chen, Lin Liu, Dunyong Zheng, Chaokui Li, and Kai Li. Supervised classification of power lines from airborne lidar data in urban areas. *Remote Sensing*, 9(8), 2017. 1, 2

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. *CoRR*, abs/1807.06521, 2018. 3

[39] Jisheng Yang, Zijun Huang, Maochun Huang, Xianxian Zeng, Dong Li, and Yun Zhang. Power line corridor lidar point cloud segmentation using convolutional neural network. In Zhouchen Lin, Liang Wang, Jian Yang, Guangming Shi, Tieniu Tan, Nanning Zheng, Xilin Chen, and Yanning Zhang, editors, *Pattern Recognition and Computer Vision*, pages 160–171, Cham, 2019. Springer International Publishing.

[40] Jianghong Zhao, Yinrui Wang, Yuee Cao, Ming Guo, Xianfeng Huang, Ruiju Zhang, Xintong Dou, Xinyu Niu, Yuanyuan Cui, and Jun Wang. The fusion strategy of 2d and 3d information based on deep learning: A review. *Remote Sensing*, 13(20), 2021.

[41] M. Zhou, K. Y. Li, J. H. Wang, C. R. Li, G. E. Teng, L. Ma, H. H. Wu, W. Li, H. J. Zhang, J. Y. Chen, and L. S. Chen. Automatic extraction of power lines from uav lidar point clouds using a novel spatial feature. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:227–234, 2019. 1

[42] S. M. Iman Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogério Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. Dublincity: Annotated lidar point cloud and its applications. *CoRR*, abs/1909.03613, 2019. 3