AttAN: Attention Adversarial Networks for 3D Point Cloud Semantic Segmentation

Gege Zhang [*], **Qinghua Ma** [*], **Licheng Jiao** [†], **Fang Liu** and **Qigong Sun**
Xidian University
{ggzhang_1, qhma}@stu.xidian.edu.cn, lchjiao@mail.xidian.edu.cn, {f63liu, xd_qigongsun}@163.com

Abstract

3D point cloud semantic segmentation has attracted wide attention with its extensive applications in autonomous driving, AR/VR, and robot sensing fields. However, in existing methods, each point in the segmentation results is predicted independently from each other. This property causes the non-contiguity of label sets in three-dimensional space and produces many noisy label points, which hinders the improvement of segmentation accuracy. To address this problem, we first extend adversarial learning to this task and propose a novel framework Attention Adversarial Networks (AttAN). With high-order correlations in label sets learned from the adversarial learning, segmentation network can predict labels closer to the real ones and correct noisy results. Moreover, we design an additive attention block for the segmentation network, which is used to automatically focus on regions critical to the segmentation task by learning the correlation between multi-scale features. Adversarial learning, which explores the underlying relationship between labels in high-dimensional space, opens up a new way in 3D point cloud semantic segmentation. Experimental results on ScanNet and S3DIS datasets show that this framework effectively improves the segmentation quality and outperforms other state-of-the-art methods.

1 Introduction

In recent years, 3D point cloud data has quickly become a research hotspot due to its abundant scene information. With extensive applications in autonomous driving, AR/VR industry and robot sensing, 3D vision tasks attract many researchers to make great efforts, e.g., 3D object detection [Shi *et al.*, 2019], 3D object classification [Thomas *et al.*, 2019] and 3D semantic segmentation [Wang *et al.*, 2019; Huang *et al.*, 2019]. Among them, 3D point cloud semantic segmentation which assigns semantic labels to points is a challenging task. Because of the unsorted and unstructured characteristics of point clouds, it is difficult to learn the feature representation between points. In earlier works, researchers first transform original point clouds into handcrafted voxel [Tchapmi *et al.*, 2017] or multi-view image [Su *et al.*, 2015] features before feeding them to Deep Convolution Neural Networks. However, these methods introduce extra computation, while also changing the raw data format.

Recently, PointNet [Qi *et al.*, 2017a] and PointNet++ [Qi *et al.*, 2017b] as pioneering works use a symmetric function and a hierarchical neural network separately to process point clouds directly. Moreover, many works [Wang *et al.*, 2019; Huang *et al.*, 2019] related to graph structure also preform well in this task. Nevertheless, a common property across all these methods is that labels are predicted independently from each other, ignoring the correlation between adjacent labels. This case results in spatial non-contiguity of label sets and produces many noisy label points. But the high-order features extracted by adversarial networks can serve as extra supervised information for the training of the main task network [Luc *et al.*, 2016]. This motivates us to further explore underlying relationships among predicted labels in high-dimensional space by adversarial learning.

In this paper, we design a framework Attention Adversarial Networks (AttAN) for 3D point cloud semantic segmentation. It consists of the segmentation network S, the Gumbel-Softmax estimator GS and the adversarial network A. During the adversarial training between S and A, high-order correlations in predicted label sets and real ones are understood by A, guiding S to predict labels following consistent distribution of the real ones [Li *et al.*, 2019]. This adversarial training process corrects noisy results, thereby effectively improves the segmentation quality. Besides, additive attention, integrating multi-scale local features for different regions, is exploited in S to improve model sensitivity. The contributions of our work are summarized as follows:

1. We first propose a novel framework Attention Adversarial Networks (AttAN), based on adversarial learning, to capture the spatial contiguity between predicted labels for 3D point cloud semantic segmentation task.
2. We propose an effective additive attention block for segmentation network to automatically focus on different regional features that are beneficial to the segmentation.

[*]Equal Contribution
[†]Corresponding Author

3. We conduct experiments on two public datasets ScanNet [Dai *et al.*, 2017] and Stanford Large-Scale 3D Indoor Spaces (S3DIS) [Armeni *et al.*, 2016]. The results demonstrate the effectiveness of our proposed approach.

2 Related Work

3D Point Cloud Segmentation. Recent works on this issue can be mainly summarized as point-based methods [Qi *et al.*, 2017b; Li *et al.*, 2018b], graph-based methods [Wang *et al.*, 2019; Huang *et al.*, 2019] and other methods [Tatarchenko *et al.*, 2018; Thomas *et al.*, 2019]. The most groundbreaking point-based works are PointNet [Qi *et al.*, 2017a] and PointNet++ [Qi *et al.*, 2017b], proposing a symmetric transformation and a hierarchical neural network to capture local features separately. Then PointCNN [Li *et al.*, 2018b] introduces X-transformation learned from point cloud itself to transform point clouds and make them satisfy certain canonical order. The above mentioned works are committed to aggregating local neighborhoods features. Besides, graph-based methods use graph structure to learn correlation in local surface patches. For example, TextureNet [Huang *et al.*, 2019] exploits high resolution signals with a new 4-RoSy surface convolution kernel on 3D surface meshes. Moreover, some other works adopt different methods for tackling orderless point clouds. For instance, Tangent Convolutions [Tatarchenko *et al.*, 2018] projects local neighborhoods to tangent planes and uses traditional 2D convolutions to process them.

Adversarial Learning. Attracted by the excellent performance of adversarial learning, the work [Luc *et al.*, 2016] uses the adversarial network in the semantic segmentation task for the first time. In text generation task, SentiGAN [Wang and Wan, 2018] combines adversarial network with reinforcement learning (RL), considering the process of generating discrete text data as a decision-making process with reward. [Chen *et al.*, 2018] applies adversarial training to generate discrete text by providing a differentiable approximation sampled from the Gumbel-Softmax estimator [Jang *et al.*, 2016]. Compared to high-variance gradient estimates existed in RL-based methods, Gumbel-Softmax is more friendly for optimizing networks with low-variance gradients, which can improve stability and speed of training [Chen *et al.*, 2018]. Previous works mostly focus on applying adversarial learning to 2D image segmentation or text generation. Additionally, PC-GAN [Li *et al.*, 2018a] applies adversarial learning to point cloud generation, but here we discuss the usage of adversarial pattern in labeling 3D point cloud datasets.

Attention Modules. Attention modules originate from human visual attention mechanism, representing the behavior that human focus on representative characteristics in images while ignoring other irrelevant information. They have been commonly applied in Natural Language Processing field. For instance, [Anderson *et al.*, 2018] designs a combined bottom-up and top-down attention mechanism, extracting features at the level of objects for image captioning and visual Q&A tasks. The architecture [Vaswani *et al.*, 2017], combining self-attention mechanism with other recurrence and convolution layers, has been successfully applied to machine translation. In image vision field, the work [Wang *et al.*, 2017] uses additive soft probabilistic attention to achieve more accurate image classification performance. The work [Oktay *et al.*, 2018] proposes a flexible attention gate model for medical image segmentation, where these gates can automatically learn to focus on targets. The work [Fu *et al.*, 2018] proposes a position attention module and a channel attention module to integrate local features with their global dependencies for scene segmentation.

3 Approach

In this paper, we propose a novel method called AttAN (illustrated in Figure 1) for segmenting raw point cloud data. Innovatively, adversarial learning is first introduced to 3D point cloud semantic segmentation, refining results by strengthening high-dimensional consistency. Specifically, the K-dimensional feature vectors produced by S are first sampled into one-hot encoding by GS. Then the fake labels and the real ones are sent to A to extract high-order features. Since they are separately considered as a whole rather than independent individuals, correlation between labels can be learned by A. The characteristics of fake labels as extra supervised information are delivered back to S and gradually become consistent with the ones of real.

In adversarial training, the discreteness of segmentation results can greatly hinder the back propagation of gradients from the adversarial network to the segmentation network, restraining the improvement of accuracy. To address this problem, the Gumbel-Softmax estimator is adopted to connect S and A. In addition, to enhance feature representation for critical areas, we recursively apply our proposed attention block in skip connections of S. Additional details are discussed in three parts below: the segmentation network, the Gumbel-Softmax estimator and the adversarial network.

3.1 The Segmentation Network

The segmentation network shown in Figure 1 adopts an encode-decode framework, following [Qi *et al.*, 2017b]. It takes point clouds as input directly and produces K-dimensional feature vectors. In the encoding stage, set abstraction (SA) modules [Qi *et al.*, 2017b] are used for hierarchical feature embedding. In the decoding stage, feature propagation (FP) modules [Qi *et al.*, 2017b] propagate features from sampled points to the origin points. Specially, one main component of S is the attention block, which is designed to learn dependency relationship between features on multiple scales in skip connection. In the following subsections, we first introduce the block, and then present the architecture of S with attention blocks .

Attention Block

Attention mechanism is usually achieved by weighted attention vectors and mainly works in three forms: hard attention mechanism, soft attention mechanism and self-attention mechanism. Among them, soft attention is probabilistic and can spread gradient to the other parts of the network. In addition, it can model the relationship between features on multiple scales and focus on useful regions. Specifically, additive attention has higher accuracy than multiplicative attention experimentally [Oktay *et al.*, 2018]. Therefore, our attention

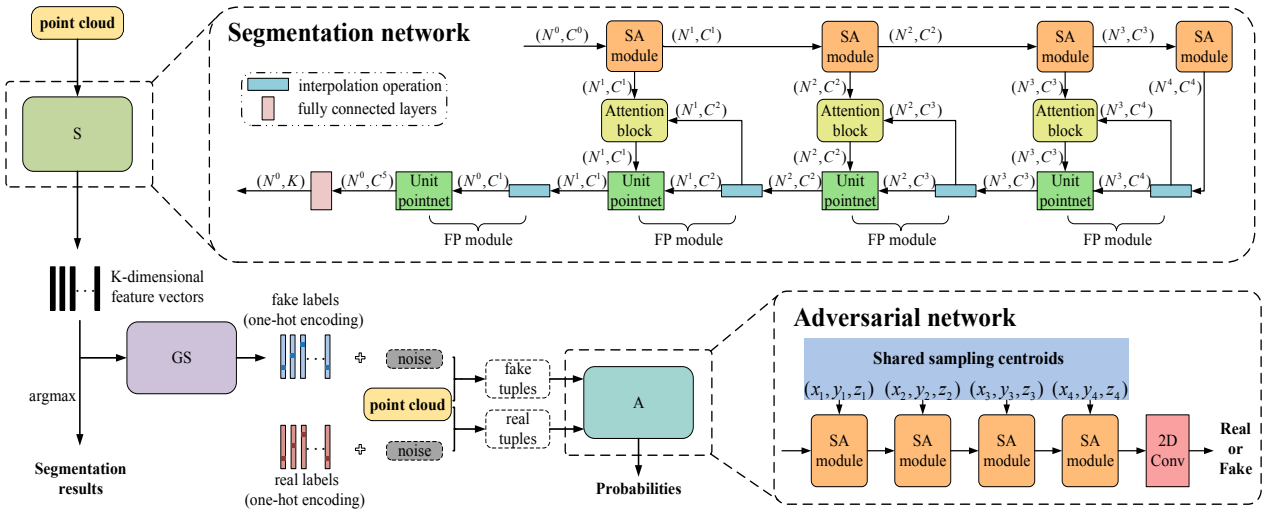

Figure 1: The overview of AttAN. It consists of three parts: the segmentation network S, the Gumbel-Softmax estimator GS and the adversarial network A.

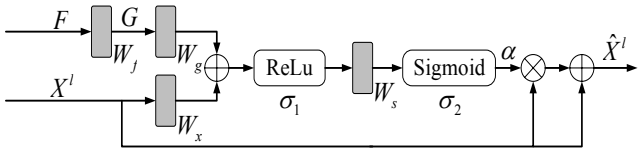

Figure 2: Schematic of the proposed attention block.

block is designed based on additive soft attention to learn critical features of objects for segmentation. As illustrated in Figure 2, note $F \in \mathbb{R}^{N^l \times C^{l+1}}$ as high level features and $X^l \in \mathbb{R}^{N^l \times C^l}$ as the output low level features of layer $l-1$ in the encoding stage, where N^l is the point set size. C^{l+1} and C^l (with $C^{l+1} > C^l$) represent the number of feature maps in different layers. We first use a 1D convolution layer $W_f \in \mathbb{R}^{C^{l+1} \times C^l}$ to reduce the dimension of F and obtain features $G \in \mathbb{R}^{N^l \times C^l}$. Next, we calculate attention masks α^l between G and X^l as follows:

$$\alpha_i^l = \sigma_2(W_s^T(\sigma_1(W_x^T x_i^l + W_g^T g_i))) \tag{1}$$

where $i = 1, ..., N^l$, σ_1 is the non-linear function $relu$, and σ_2 corresponds to the $sigmoid$ activation function. $W_x \in \mathbb{R}^{C^l \times C^l}$, $W_g \in \mathbb{R}^{C^l \times C^l}$ and $W_s \in \mathbb{R}^{C^l \times C^l}$ represent 1D convolution layers. The attention masks α indicate the significance of different regions. Then we perform element-wise multiply operation between α^l and X^l. Finally, the residual connection similar to [Wang *et al.*, 2017] is employed to retain original features, so the final output $\hat{X}^l$ of attention block is defined as:

$$\hat{x}_i^l = \alpha_i^l x_i^l + x_i^l \tag{2}$$

The attention block can act like a feature selector to augment the useful structure features automatically during forward process by replacing X^l with the weighted attention features $\hat{X}^l$. Additionally, it is a gradient update filter during back propagation. The update rule for convolution parameters ϕ^{l-1} in layer $l-1$ is:

$$\begin{aligned} \frac{\partial \hat{X}^l}{\partial \phi^{l-1}} &= \frac{\partial(\alpha^l f(X^{l-1}; \phi^{l-1}) + f(X^{l-1}; \phi^{l-1}))}{\partial \phi^{l-1}} \\ &= (\alpha^l + 1)\frac{\partial f(X^{l-1}; \phi^{l-1})}{\partial \phi^{l-1}} + \frac{\partial \alpha^l}{\partial \phi^{l-1}} X^l \end{aligned} \tag{3}$$

Such a block allows the neural network to highlight salient regional features passing through the skip connections and can achieve better segmentation results.

The Architecture of Segmentation Network with Attention Blocks

SA and FP modules are two significant parts proposed in PointNet++ [Qi *et al.*, 2017b]. The SA module in layer $l-1$ takes point cloud data with shape $N^{l-1} \times C^{l-1}$ as input. By means of the farthest point sampling, grouping as well as a unit pointnet[1], it generates output feature embedding vectors of sampling centroids with shape $N^l \times C^l$, which will be sent to the attention block as X^l. To get point-wise prediction results, the FP module first interpolates inverse distance weighted feature values of the input N^{l+1} points at the coordinates of N^l points based on k-nearest neighbors and generates interpolated features of size $N^l \times C^{l+1}$ (with $N^l > N^{l+1}$). These features are sent to our proposed attention block as F, to calculate the attention masks with X^l. Then the output weighted attention features $\hat{X}^l$ concatenated with F are sent to a unit pointnet. The process is recursively performed until we have features with same size as the original point set. Then fully connected layers are used for segmentation. The attention blocks can aggregate information from multiple scales and eliminate responses of irrelevant and noisy parts in skip connections, which contributes to the segmentation task.

[1] An architecture defined in PointNet [Qi *et al.*, 2017a].

3.2 The Gumbel-Softmax Estimator

The exploration of adversarial learning is not straightforward, owing to the problems interpreted next. Note $o_s \in \mathbb{R}^{N\times K}$ as the output feature vectors of S , we calculate the segmentation results $l_s = argmax(o_s) \in \mathbb{R}^{N\times 1}$. Assuming l_s are sent into A for training, the standard backward propagation of gradients from A to S cannot be applied since the non-differentiable $argmax$ function. Besides, numerical values of l_s also implicitly increase the between-class distance.

Alternately, the differentiable probability values v_s are generated by $softmax$[2]: $v_s = softmax(o_s) \in \mathbb{R}^{N\times K}$. They can approximate one-hot distribution by enlarging the difference between elements of o_s and maintain original between-class distance. However, they can be distinguished by A from the one-hot encoding of ground truth labels $y_r \in \mathbb{R}^{N\times K}$, because these two distributions have no overlap and the training samples based on $softmax$ lack of randomness.

To address the above obstacles, we apply a straight-through Gumbel-Softmax estimator [Jang *et al.*, 2016] to sample the output o_s of S:

$$y_s = softmax((o_s + g)/\tau) \tag{4a}$$

$$y_f = onehot(argmax(y_s)) \tag{4b}$$

where g represents the Gumbel noise sampled from the i.i.d standard Gumbel distribution [Jang *et al.*, 2016], $\tau \in (0, \infty)$ is the temperature , $y_s \in \mathbb{R}^{N\times K}$ is re-parameterized probability of o_s and $y_f \in \mathbb{R}^{N\times K}$ is one-hot encoding of y_s. With the decrease of temperature, Gumbel-Softmax distribution can be smoothly annealed into categorical distribution. However, straight-through Gumbel-Softmax estimator allows samples to be sparse even when the temperature is high [Jang *et al.*, 2016]. In our experiments, τ is set to 1.0. Due to the "$onehot$ with $argmax$" in Eq.4b is non-differentiable, we use continuous probability in the backward pass by approximating:

$$\nabla_\theta y_s \approx \nabla_\theta y_f \tag{5}$$

The Gumbel-Softmax estimator makes probabilities meaningful since it can randomly generate some non-maximum probability category samples according to its random noise, which is essential for training models in adversarial pattern.

3.3 The Adversarial Network

In this work, the aim of adversarial network is to distinguish the inconsistencies in high-dimensional space between the inputs y_r and y_f for label refining. As illustrated in Figure 1, we add noise to the inputs artificially for stabilizing training [Arjovsky *et al.*, 2017]. Then the new real input encoding i_r and fake input encoding i_f will concatenate with point clouds as real and fake input tuples of A separately. The overall structure of A consists of four SA modules and a 2D convolution layer. As an important cue, A can extract high-dimensional features of input tuples and generate the probabilities that input tuples are real.

Specially, the SA modules can be used in A, due to i_r and i_f both represent extra information of point cloud. However, we remove the batch normalization layers of SA modules to train A effectively. It is worth mentioning that the coordinates of same sampling centroids are shared by S and A for providing sampling related information. Getting through the architecture of A, y_f and its corresponding ground truth y_r have comparative characteristics in high-dimensional space. The adversarial network can detect mismatches between characteristics flexibly without manual work. At the same time, it considers labels as a whole and learns the correlation between labels, which is utilized in the training of S. In this way, adversarial training provides strong regularization for deep networks and revises unreasonable segmentation results.

[2] $softmax(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{K} exp(x_j)}, i = 1, ..., K$

Algorithm 1 The adversarial training process in AttAN

Input: Input point clouds p; The number of batches M; Learning rate η; Maximum epoch E
Parameter: θ, φ
Output: Well trained S_θ

```
Initialize S_θ, A_φ with random weights θ, φ.
Pre-train S_θ by minimizing the first term of Eq.6.
for i = 0 to E − 1 do
    for j = 0 to M − 1 do
        Sample a minibatch of point clouds p^j and the corresponding ground truth labels y_r^j.
        Obtain the output feature vectors of S_θ:
                        o_s^j = S_θ(p^j)
        Apply Gumbel-Softmax estimator to sample o_s^j and obtain y_f^j using the Eq.4.
        Add noise to both real labels y_r^j and fake labels y_f^j.
        if i is Even then
            Update φ by minimizing the Eq.7.
        end if
        Update θ by minimizing the Eq.6.
    end for
end for
return well trained S_θ
```

4 Objectives and Optimization

In this section, we introduce the objective functions for segmentation network and adversarial network. In addition, the optimization algorithm is described in detail.

Objective for Segmentation Network. The objective of segmentation network is defined as:

$$\mathcal{L}_S = \mathcal{L}_{cross}(o_s, y_r) + \mathcal{L}_{mse}(A(p, i_f), 1) \tag{6}$$

where p represents the input point cloud. We use multi-class cross entropy loss to encourage the segmentation network to predict the right labels. In addition, mean square error (MSE) function, as the adversarial part, provides a stronger gradient term when A makes accurate predictions on input tuples. So that it can push the segmentation network to predict labels which are hard to distinguish from the ground truth ones for the adversarial network. Training the segmentation network is equal to minimizing the above objective function.

Objective for Adversarial Network. The adversarial network learns to discriminate the real and the fake tuples using

the following objective function:

$$\mathcal{L}_A = \mathcal{L}_{mse}(A(p, i_f), 0) + \mathcal{L}_{mse}(A(p, i_r), 1) \quad (7)$$

Both terms use MSE as loss function. We optimize the adversarial network by minimizing the above objective function. The adversarial training is reflected in the second term of Eq.6 and the first term of Eq.7.

We optimize the proposed networks in the following approach: Firstly, to provide a good initialization of S for a good convergence behavior in adversarial training, we pretrain S by minimizing the first term of Eq.6. Then training process is executed in way of training A on every even epoch and S on every epoch, which can reduce the learning speed of A to avoid overfitting and give S certain space to grow. The detailed algorithm is described in Algorithm 1.

We use an Adam optimizer with momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$, and a batch size of 16 to train the model. In addition, following prior works [Qi *et al.*, 2017a; Qi *et al.*, 2017b], we apply exponential decay learning rate policy to adjust learning rate in every batch. The base learning rate is set to 0.001. Specially, we employ exponential moving average strategy with decay rate of 0.99 to reduce fluctuations in the training procedure. At inference time, we only run the segmentation network in exactly the same way as during the training phase. After getting o_s, we use $argmax$ function to obtain the final segmentation results.

5 Experiments

To evaluate our approach, we conduct extensive and comprehensive experiments on public semantic segmentation datasets ScanNet [Dai *et al.*, 2017] and S3DIS [Armeni *et al.*, 2016]. Experimental results show that AttAN achieves state-of-the-art performance. In the following subsections, we first introduce these datasets, and then analyze a series of ablation experimental results to demonstrate the effectiveness of our proposed architecture on ScanNet dataset. Finally, we show segmentation results on S3DIS dataset.

5.1 Datasets

ScanNet. The newest version of this dataset includes 1513 scanned and reconstructed scenes with 21 semantic classes and 100 new test scenes with all semantic labels publicly unavailable. During the training phase, we use 1201 scenes for training and 312 scenes for validating, both without extra RGB information. Then we submit our results on test scenes to the official benchmark evaluation server[3] to compare against other methods. In this dataset, we use the mean of intersection over union (mIoU) across all the categories as evaluation metric, which is same as the benchmark.

Stanford Large-Scale 3D Indoor Spaces (S3DIS). This dataset contains scanned point cloud data of 271 rooms in 6 areas. Each point in the point cloud sets is assigned a label from 13 categories. We process the dataset similar as [Qi *et al.*, 2017a]. Firstly, we split points by rooms, and then sample rooms into blocks with area 1m by 1m. In the training stage we randomly sample 4096 points in each block while in the testing stage we test on all points. Area 5 with some different objects is not the same building as other areas, which as test set could better measure the generalization ability of our method. Moreover, following [Armeni *et al.*, 2016; Qi *et al.*, 2017a], 6-fold cross validation on all areas is adopted for further evaluation.

[3]http://kaldir.vc.in.tum.de/scannet_benchmark/semantic_label_3d

Method	Gumbel-Softmax	Add Noise	mIoU(%)
PointNet++			49.89
PANet	√		50.20
PANet-noise	√	√	**50.70**

Table 1: Ablation study on ScanNet for adversarial learning.

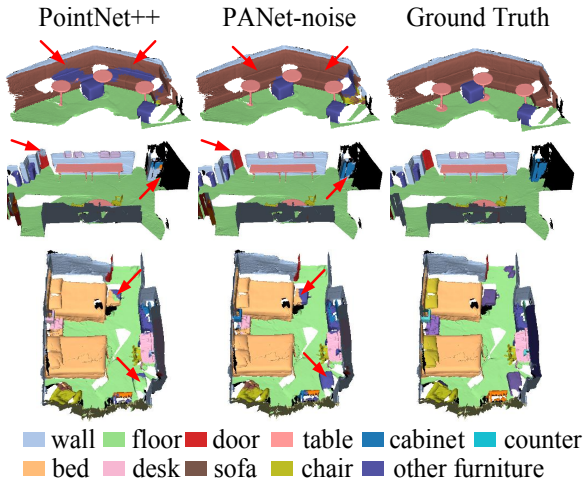

Figure 3: The segmentation results using different networks. Left: The segmentation results of PointNet++. Middle: The segmentation results of PANet-noise. Right: Ground Truth.

5.2 Results on ScanNet Dataset

Ablation Study for Adversarial Learning. To verify the effectiveness of adversarial learning, we design a framework named PointNet++ Adversarial Networks (PANet). It combines the segmentation network PointNet++ [Qi *et al.*, 2017b] with our Gumbel-Softmax estimator and adversarial network. Then we conduct experiments on ScanNet [Dai *et al.*, 2017] validation scenes with different strategies in Table 1. We can see that PANet improves mIoU by about 0.3% against baseline PointNet++, which proves that the adversarial pattern can be extended in 3D point cloud semantic segmentation task with the Gumbel-Softmax estimator. In addition, PANet-noise (with noise added to both real one-hot encoding and fake one-hot encoding of PANet) raises the mIoU continuously by 0.5%. This demonstrates that it is necessary and effective to add noise in adversarial training process. The segmentation results on some validation scenes using PointNet++ and PANet-noise are visualized in Figure 3. The misclassified and unreasonable labels pointed by red arrows can be correctly classified using adversarial learning.

Method	Attention Blocks	mIoU(%)
PointNet++		49.89
AttPointNet++	√	50.12
PANet-noise		50.70
AttAN	√	**51.18**

Table 2: Ablation study on ScanNet for attention blocks. AttPointNet++ represents PointNet++ with attention blocks.

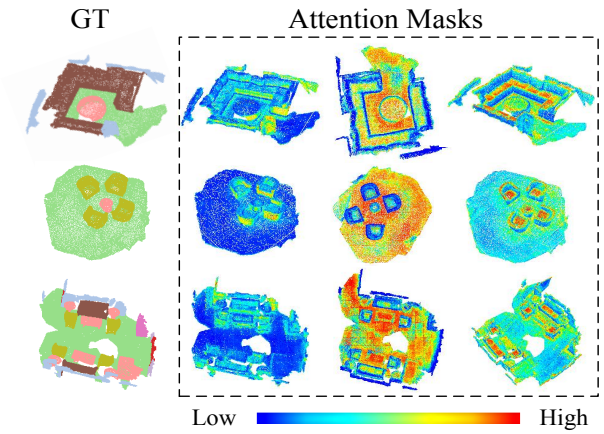

Figure 4: The visualization of attention masks in last attention block. Left: Ground Truth. Right: Attention masks of different feature channels. The color from blue to red indicates different weight values for each point feature.

Ablation Study for Attention Blocks. We design attention blocks to capture more useful information during training stage. To verify the effectiveness of attention blocks, we conduct ablation experiments with different settings on ScanNet [Dai *et al.*, 2017] validation scenes. The segmentation results are described in Table 2. We can observe that the attention blocks improve the performance remarkably. Compared with the baseline PointNet++, the model AttPointNet++ can bring nearly 0.25% improvement in mIoU. We also employ attention blocks in PANet-noise mentioned above and get the model AttAN. These blocks result in almost 0.5% promotion in mIoU. Furthermore, to better analyze the effects of attention blocks, we visualize attention masks in last attention block in Figure 4. It can be seen that the attention masks in different feature channels focus on diverse regions or geometry structures, which are of great help for the semantic segmentation. That means attention blocks can learn to assign low weights for unimportant point features and high weights for more discriminative point features. Concretely, the first column of attention masks indicates that vertical outlines are enhanced. The second column pays more attention to the horizontal planes of objects. From the third column, we can observe sofas, tables and chairs obviously, which indicates the masks contribute to segmenting these categories.

Comparison with State-of-the-art. We compare our method against other existing methods and results are shown in Table 3. All other methods use both color and geometry

Method	mIoU(%)
PointNet++ [Qi *et al.*, 2017b]	33.9
SPLAT Net [Su *et al.*, 2018]	39.3
Tangent Convolutions [Tatarchenko *et al.*, 2018]	43.8
3DMV [Dai and Nießner, 2018]	48.4
TextureNet [Huang *et al.*, 2019]	56.6
AttAN	**60.9**

Table 3: Comparison with the state-of-the-art methods for 3D semantic segmentation on the ScanNet test scenes.

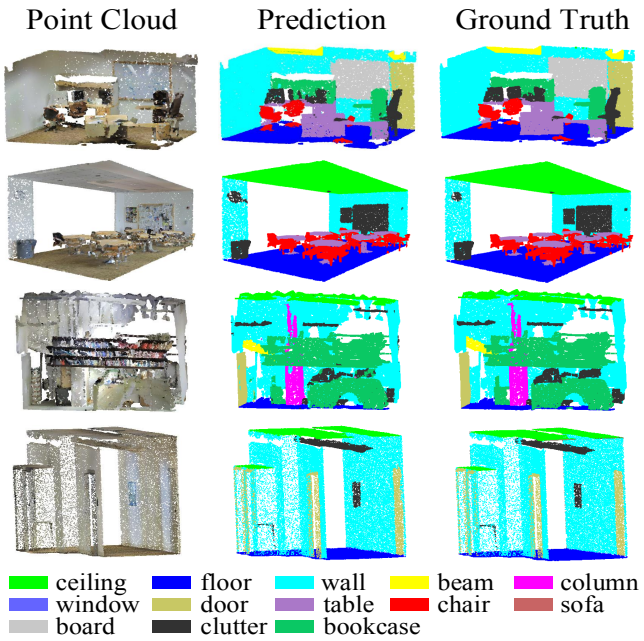

Figure 5: The visualization results on Area 5 of S3DIS.

in naive formats of 3D point cloud and report their results on ScanNet test scenes to the benchmark. In our experiments, we use 3D coordinates data without extra RGB information. AttAN outperforms these state-of-the-art methods by a significant margin. Specifically, the mIoU of ours is 27% higher than baseline PointNet++ (33.9% vs. 60.9%), and it exceeds the previous state-of-the-art method TextureNet (56.6% vs. 60.9%) by around 4%.

5.3 Results on S3DIS Dataset

In this subsection, we conduct experiments on S3DIS dataset to further evaluate the effectiveness of our proposed method. First, comparison with previous state-of-the-art methods on Area 5 are showed in Table 4. AttAN outperforms other methods with dominant advantage, achieving the best OA and mIoU. Concretely, 7 of 13 categories achieve the best performance. For "beam" with few points (0.029%), other state-of-the-art methods cannot perform well compared to AttAN, which also demonstrates the robustness of our method. Additional qualitative segmentation results on Area 5 are visualized in Figure 5. The rooms from top to bottom represent office, lobby, storage and hallway respectively. Among them,

Method	OA	mIoU	ceiling	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	clutter
PointNet [Qi *et al.*, 2017a]	-	41.09	88.80	97.33	69.80	0.05	3.92	46.26	10.76	58.93	52.61	5.85	40.28	26.38	33.22
SegCloud [Tchapmi *et al.*, 2017]	-	48.92	90.06	96.05	69.86	0.00	18.37	38.35	23.12	75.89	70.40	58.42	40.88	12.96	41.60
PointCNN [Li *et al.*, 2018b]	85.91	57.26	92.31	98.24	79.41	0.00	17.60	22.77	62.09	74.39	80.59	31.67	66.67	62.05	56.74
HPEIN [Jiang *et al.*, 2019]	87.18	61.85	91.47	98.16	81.38	0.00	23.34	65.30	40.02	87.70	75.46	67.78	58.45	65.61	49.36
GACNet [Wang *et al.*, 2019]	87.79	62.85	92.28	**98.27**	81.90	0.00	20.35	59.07	40.85	78.54	85.80	61.70	70.75	**74.66**	52.82
KPConv [Thomas *et al.*, 2019]	-	67.1	92.8	97.3	**82.4**	0.00	23.9	58.0	69.0	**91.0**	81.5	**75.3**	**75.4**	66.7	58.9
AttAN	**90.51**	**74.68**	**95.03**	97.24	81.74	**87.36**	**66.13**	**85.49**	**76.09**	73.28	**92.18**	63.34	56.28	29.49	**67.28**

Table 4: The comparisons on the S3DIS Area 5 in overall accuracy (OA, %), mIoU (%), and per-class IoU (%).

Method	OA(%)	mIoU(%)
PointNet [Qi *et al.*, 2017a]	78.62	47.71
SPGraph [Landrieu and Simonovsky, 2018]	85.5	62.1
PointCNN [Li *et al.*, 2018b]	88.14	65.39
HPEIN [Jiang *et al.*, 2019]	88.20	67.83
KPConv [Thomas *et al.*, 2019]	-	70.6
AttAN	**90.26**	**72.39**

Table 5: Comparison with the state-of-the-art methods on the S3DIS dataset with 6-fold cross validation.

the office room have most "beam" points of the whole area that can be predicted decently by AttAN. Although several noisy points haven't been corrected by the adversarial learning yet, the regular geometry structure of different objects are recognized well, which also proves the advantage of attention blocks in 3D point cloud segmentation. Table 5 provides the comparison among different methods with 6-fold cross validation. Our method still reaches the best performance on two metrics.

6 Conclusion

In this paper, we propose a novel framework AttAN for 3D point cloud semantic segmentation. It is endowed with two key properties: First, the adversarial learning leads the segmentation network to take the correlations between labels into consideration, and thus corrects the segmentation results. Second, the attention blocks in the segmentation network automatically capture information of salient positions, giving more useful details for the segmentation task. To demonstrate the superiority of our proposed framework, we conduct quantitative experiments on two public datasets, ScanNet and S3DIS. Results show that AttAN outperforms other state-of-the-art methods. We also investigate the effectiveness of adversarial learning and attention blocks independently by ablation experiments. Specially, the Gumbel-Softmax estimator is applied to introduce adversarial learning into 3D point cloud semantic segmentation. Our method provides promising results for future work. The application of more and better estimators is still an open question worth researching.

Acknowledgements

This work was partially supported by the State Key Program of National Natural Science of China (No. 61836009), Project supported the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 61621005), the National Natural Science Foundation of China (Nos. U1701267, 61871310, 61773304, 61806154, 61802295 and 61801351), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No. B07048), the Major Research Plan of the National Natural Science Foundation of China (Nos. 91438201 and 91438103), the Program for Cheung Kong Scholars and Innovative Research Team in University (No. IRT_15R53).

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[Armeni *et al.*, 2016] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016.

[Chen *et al.*, 2018] Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pages 4666–4677, 2018.

[Dai and Nießner, 2018] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018.

[Dai *et al.*, 2017] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[Fu *et al.*, 2018] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018.

[Huang *et al.*, 2019] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J

Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4440–4449, 2019.

[Jang *et al.*, 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[Jiang *et al.*, 2019] Li Jiang, Hengshuang Zhao, Shu Liu, Xiaoyong Shen, Chi-Wing Fu, and Jiaya Jia. Hierarchical point-edge interaction network for point cloud semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10432–10440, 2019.

[Landrieu and Simonovsky, 2018] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2018.

[Li *et al.*, 2018a] Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018.

[Li *et al.*, 2018b] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018.

[Li *et al.*, 2019] Jing Li, Deheng Ye, and Shuo Shang. Adversarial transfer for named entity boundary detection with pointer networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5053–5059. AAAI Press, 2019.

[Luc *et al.*, 2016] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017.

[Shi *et al.*, 2019] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

[Su *et al.*, 2015] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

[Su *et al.*, 2018] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2018.

[Tatarchenko *et al.*, 2018] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2018.

[Tchapmi *et al.*, 2017] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 International Conference on 3D Vision (3DV)*, pages 537–547. IEEE, 2017.

[Thomas *et al.*, 2019] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. *arXiv preprint arXiv:1904.08889*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang and Wan, 2018] Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452, 2018.

[Wang *et al.*, 2017] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[Wang *et al.*, 2019] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10296–10305, 2019.