diffConv: Analyzing Irregular Point Clouds with an Irregular View (Supplemental Material)

Manxi Lin and Aasa Feragen

Technical University of Denmark, Kongens Lyngby, Denmark
{manli,afhar}@dtu.dk

1 Network Architecture

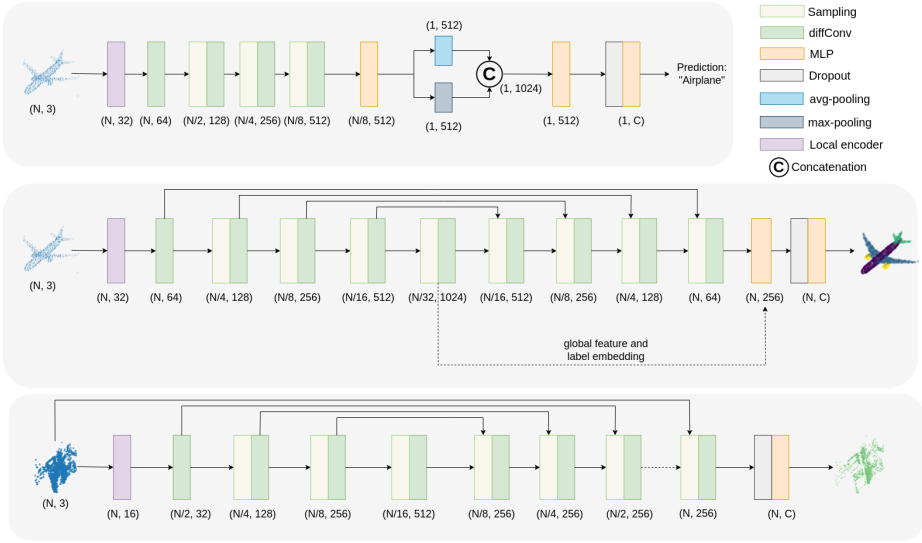


Fig. 1. Overview network architectures for classification **(top)** 3D shape part segmentation **(middle)**, and scene semantic segmentation **(bottom)**. Specifically, in the encoders, 'Sampling' denotes random downsampling, while in the decoders, it denotes feature interpolation (point upsampling).

Fig. 1 shows our network for different tasks, where diffConv is performed hierarchically to capture multi-scale point features and avoid redundant computation.

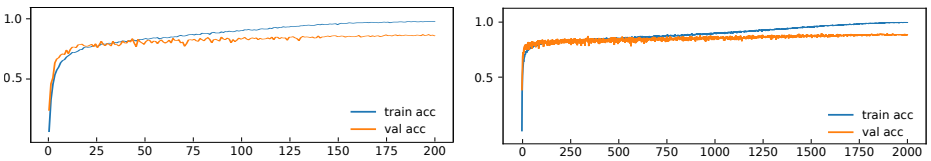
Specifically, input point coordinates are initially encoded to a higher dimension by a local feature encoder. For classification, the encoder is an MLP, for segmentation, the encoder is a point abstraction module [3]. Then, point features are grouped and aggregated through multiple diffConvs. In contrast to the

widely-adopted furthest point sampling [3], during encoding, we select key points by random sampling, which has recently been proved efficient and effective [2]. For classification, we follow the global aggregation scheme of DGCNN [5], where the learned local features are pooled by a max-pooling and an average-pooling (avg-pooling) respectively. The pooled features are concatenated and processed by MLPs. For segmentation, we use the same attention U-Net style decoder architecture with CurveNet. In the 3D shape segmentation task, we fuse the global feature and the label embedding of object shape category with the learned features, following DGCNN [5].

2 Training Details

Settings. For all experiments, the squared initial searching radius r^2 was set to 0.005, and increased to the inverse of the sampling rate times after each sampling; the kernel density bandwidth h was set to 0.1. Gaussian error linear units [1] were used as nonlinear activation. In 3D object classification and part segmentation, we trained the models by optimizing cross-entropy loss with label smoothing, using SGD with a learning rate of 0.1 and a momentum of 0.9, using batch size 32 in training and 16 in testing. In the scene segmentation task, the optimizer was AdamW with a learning rate of 0.001 and the training batch size was 16. The learning rate was reduced by cosine annealing to 0.001. The random seed was fixed at 42 in all experiments to enhance reproducibility. The dropout rate was set to 0.5, and we applied random scaling within $[\frac{2}{3}, \frac{3}{2}]$, random translation within $[-0.2, 0.2]$ and shuffling as augmentation. In scene segmentation, we also jittered the points within the range of ± 0.01 during training. We trained the model for 2,000 epochs in all the tasks.

Training process. Our model is trained for more epochs for full convergence, compared to existing methods. Fig. 2 shows the learning curves of our model and CurveNet, and our training is as stable as CurveNet. The robustness study on ModelNet40-C also proves our model is not overfitting to the dataset. Note that this training setting does not affect our model's faster inference speed.



(a) CurveNet training curve (b) diffConv training curve

Fig. 2. Training curves of our model and CurveNet in ModelNet40 classification task.

3 Ablation Studies

Ablation study over components. To verify the effectiveness of the proposed diffConv, we conducted a detailed ablation study over components on the ModelNet40 official split. The experimental settings and data augmentation was the same with the ModelNet40 classification task.

Table 1. Ablation study over components. "LS" refers to Laplacian smoothing, "MAT" denotes masked attention, "DDBQ" stands for density-dilated ball query, and "BR" denotes balanced renormalization. Note that the equations refer to the equations in the main paper.

LS	DDBQ	MAT	BR	OA(%)	MA(%)
Eq. 3	Eq. 8	Eq. 9	Eq. 11		
				89.8	85.2
✓				90.8	86.4
	✓			90.6	86.1
		✓	✓	92.7	89.9
✓	✓			92.2	89.5
✓		✓	✓	93.1	90.4
	✓	✓	✓	92.5	89.2
✓	✓	✓		92.9	89.8
✓	✓	✓	✓	**93.6**	**90.6**

We assessed the effect of employing the three main components of diffConv, namely Laplacian smoothing, density-dilated ball query, and masked attention. When Laplacian smoothing was not applied, the feature vector S in Eq. 3 in the main paper was represented as $S = \hat{A}X$. When the density-dilated ball query was removed, the model grouped points by vanilla ball query with of a radius of $\sqrt{0.005}$. When masked attention was disabled, the adjacency matrix was replaced with the binary matrix similar to Eq. 5 in the main paper. We also evaluated the effect of the balanced renormalization strategy applied in masked attention. This ablation study was done by simply replacing the balanced renormalization with the normalization strategy employed by the original self-attention [4].

Table 1 reports the overall and mean-class accuracy under different component combinations. When the model is not applying diffConv, the overall accuracy is only 89.8%. We consistently see that the three components all bring improvements to the results. In contrast to the constant-radius ball query, our density-dilated modification improves the OA by 0.5. This is in accordance with our analysis in Section 3.3 in the main paper that the long-range information flow is boosted by the dilated neighborhood. We also notice that masked attention plays a key role in diffConv. Comparing the fifth and the last row, masked attention improves the results by 1.4 and 1.1 in OA and MA separately. Besides, the employed balanced renormalization strategy shows significant improvement

on model performance, compared to the normalization adopted by the original self-attention.

Component contribution to robustness. Our model shows strong robustness on the ModelNet40-C benchmark. We studied the contribution of the proposed density-dilated ball query and masked attention to the model corruption robustness. Specifically, when masked attention was disabled, the adjacency matrix was replaced with the binary matrix as in the ablation study over components. When density-dilated ball query was disabled, we employed the KNN grouping, which is sensitive to noise points according to our analysis in Section 3.3 in the main paper. The models were trained on the ModelNet40 dataset and evaluated on the ModelNet40 benchmark by OA and MA, and the ModelNet40-C benchmark by the corruption error rate. Table 2 illustrates the experiment results.

Table 2. Component contribution to robustness. Here, "w/o DDBQ" denotes the model grouping 20 nearest neighbors instead of density-dilated ball query; "w/o MAT" stands for the model without masked attention; "Complete model" denotes the model equipped with all the proposed components.

Model types	CER(%)	OA(%)	MA(%)
w/o DDBQ	25.2	92.7	89.2
w/o MAT	21.9	92.2	89.5
Complete model	**21.4**	**93.6**	**90.6**

The results show that both the proposed density-dilated ball query and masked attention, especially the irregular ball query, contribute to the model's robustness. This supports our hypothesis of the robustness of the irregular point representation.

Attention v.s. inductive bias. According to Table 1, masked attention, which assigns each neighbor a weight, is the most effective part of diffConv. The weight does not rely on an inductive bias and is purely learned from point features as well as coordinates and updated dynamically during training. This is different from the predefined rules for point weighting applied in previous work. We compare masked attention with several intuitive inductive biases in Table 3. The models were trained and evaluated on the ModelNet40 official split. The isotropic bias [3] treats all the neighbors equally. Spatial distance [8] and feature distance bias [7] assign larger importance to the neighbor closer to the key point in the Euclidean and feature space respectively. The inverse density bias is taken from PointConv [6], which posits high-density neighbors a lower contribution. We implemented the last three biases via replacing the adjacency matrix from Eq. 9 in the main paper with the respective metrics processed by a Gaussian kernel, similar to [7].

Table 3. Results of study over attention v.s. inductive bias.

Aggregation rules	OA(%)	MA(%)
Isotropic bias	92.2 (1.4 ↓)	89.5 (1.1 ↓)
Spatial distance	90.2 (3.4 ↓)	84.9 (5.7 ↓)
Feature distance	91.1 (2.5 ↓)	86.8 (3.8 ↓)
Inverse density	89.7 (3.9 ↓)	85.0 (5.6 ↓)
Ours	**93.6**	**90.6**

Our method outperforms all the conventional inductive biases by more than 1.4 and 1.1 in OA and MA. With the introduction of pre-defined neighbor preference (rows 2, 3, and 4), the model performance becomes even worse than the isotropic bias (row 1) that treats all the neighbors evenly. In contrast to prior knowledge, the irregularity given by the density-dilated view and masked attention better exploits latent point local structure. We attribute the improvement to the introduction of irregularity.

Impact of bandwidth in kernel density estimation. Table 4 presents the impact of bandwidth h in kernel density estimation (Eq. 7 in the main paper) on ModelNet40 classification. According to the table, 0.1 is the optimal bandwidth.

Table 4. Results of our model with different kernel density bandwidths.

Kernel density bandwidth (h)	OA(%)	MA(%)
0.05	93.1	90.1
0.1	**93.6**	**90.6**
0.5	93.3	90.1

Impact of squared initial searching radius. The impact of various squared initial searching radius r^2 settings is illustrated in Table 5. According to Eq. 8 in the main paper, r^2 determines the lower bound of the searching radius. All the experiments were conducted on ModelNet40.

In line with the results, we find that with a large r^2, the model performance degenerates, since the model fails to capture point local geometric structures.

Ablation study over dilating strategies. We compared three different strategies for dilating searching radius in Eq. 8 in the main paper. Given r the pre-set initial searching radius, in strategy one, the dilated radius $r_i = r(1 + \hat{d}_i)$ is linearly correlated with the point kernel density $\hat{d}_i$. In strategy two, the squared dilated radius $r_i^2 = r^2(1 + \hat{d}_i)$ is linearly correlated with the point kernel density. In

Table 5. Results of our model with different squared initial searching radius.

Squared initial searching radius (r^2)	OA(%)	MA(%)
0.001	92.9	89.8
0.005	**93.6**	**90.6**
0.01	93.2	90.4
0.05	93.0	90.1
0.1	91.8	88.1

the last strategy, the dilated radius $r_i = r \cdot (1 + \frac{e^{\hat{d}_i} - 1}{e-1})$ has a nonlinear relationship with the kernel density. All the experiments were run on the ModelNet40 benchmark.

Table 6. Results of study over different radius dilating strategies. "Linear-1", "Linear-2" and "Exponent" denote the three strategies in the text respectively.

Strategies	OA(%)	MA(%)
Linear-1	92.6	89.6
Linear-2	**93.6**	**90.6**
Exponent	93.0	89.9

Table 6 presents the results. The second strategy, "Linear-2", achieves the best performance. This demonstrates that a too-fast dilation speed fails to benefit the point feature learning.

4 Additional Visualizations of Masked Attention

Fig. 4 in the main paper illustrates the attention maps of 8 objects from different categories on the ModelNet40-C benchmark. We visualize the attention maps of objects from the rest 32 categories in Fig. 3. The attention scores were taken from the second diffConv of the classification network. All the objects were corrupted by "the most severe background noise" (a "severity" of 5). For each object, we randomly picked two key points and visualized their neighbors. As shown in the figure, our diffConv isolates the noise points, endows the flat-area points with a larger receptive field and focuses on the neighbors with larger differences in geometric features to the key points.

We also illustrate how neighbors are selected when two flat surface approach each other, with the ground truth mask and attention score of an example point cloud from Toronto3D in Fig. 4. The red point is a building point that lies on the boundary of the building (green points) and the road (milky points). The

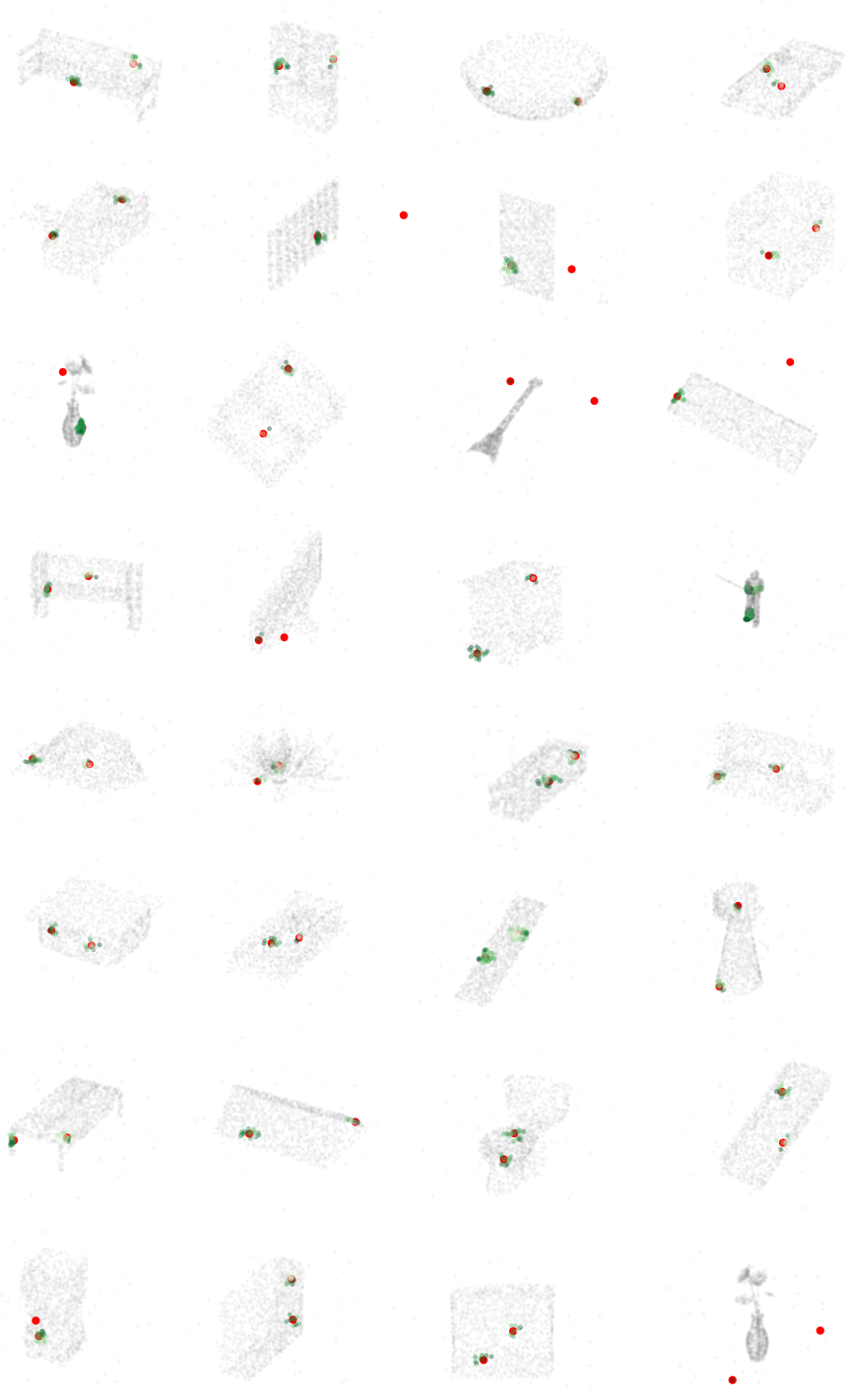


Fig. 3. Masked attention maps for ModelNet40-C. Red dots are picked key points; green dots denote neighbors; the color scale denotes the masked attention score. Zoom in for details.

building and the road are approximately flat surfaces. According to the figure, our diffConv emphasizes the neighbors from the building (with darker color).

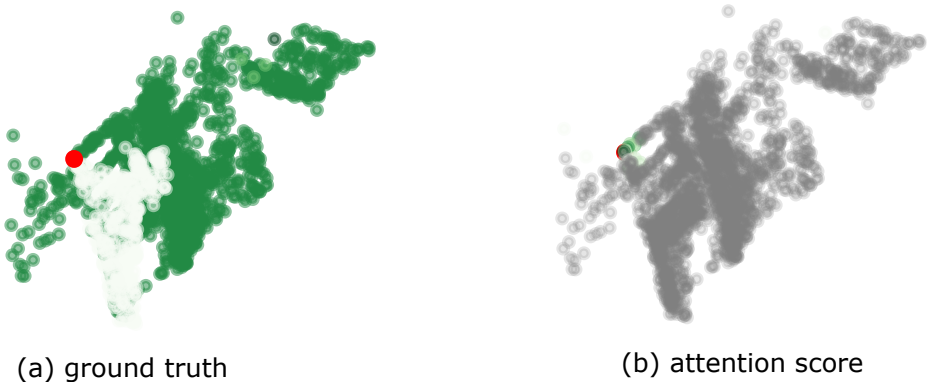


Fig. 4. Ground truth mask and attention score of an example from Toronto3D. Zoom in for details.

References

1. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
2. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11108–11117 (2020)
3. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 5105–5114 (2017)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
5. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) **38**(5), 1–12 (2019)
6. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9621–9630 (2019)
7. Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., Qiao, Y.: Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. arXiv preprint arXiv:2012.10921 (2020)
8. Yi, L., Su, H., Guo, X., Guibas, L.J.: Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2282–2290 (2017)