ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097060

D-CONFORMER: DEFORMABLE SPARSE TRANSFORMER AUGMENTED CONVOLUTION FOR VOXEL-BASED 3D OBJECT DETECTION

Xiao Zhao[1], *Liuzhen Su*[1], *Xukun Zhang*[1], *Dingkang Yang*[1], *Mingyang Sun*[1], *Shunli Wang*[1], *Peng Zhai*[1], *Lihua Zhang*[†1,2]

[1]Academy for Engineering and Technology, Fudan University, Shanghai, China
[2]Engineering Research Center of AI and Robotics, Shanghai, China

ABSTRACT

Although CNN-based and Transformer-based detectors have made impressive improvements in 3D object detection, these two network paradigms suffer from the interference of insufficient receptive field and local detail weakening, which significantly limits the feature extraction performance of the backbone. In this paper, we propose to fuse convolution and transformer, and simultaneously considering the different contributions of non-empty voxels at different positions in 3D space to object detection, it is not consistent with applying standard convolution and transformer directly on voxels. Specifically, we design a novel deformable sparse transformer to perform long-range information interaction on fine-grained local detail semantics aggregated by focal sparse convolution, termed D-Conformer. D-Conformer learns valuable voxels with position-wise in sparse space and can be applied to most voxel-based detectors as a backbone. Extensive experiments demonstrate that our method achieves satisfactory detection results and outperforms state-of-the-art 3D detection methods by a large margin.

***Index Terms*—** 3D object detection, deformable sparse transformer, focal sparse convolution, KITTI dataset

1. INTRODUCTION

Point cloud-based 3D object detection is widely used in autonomous vehicles. Extracting features from irregular and sparse point clouds is the current challenge. Voxel-based methods [1, 2, 3] convert raw point clouds into regular voxel grids and process them using 3D sparse convolutional neural networks (Sparse CNNs) [1]. Sparse CNNs consist of regular sparse convolution and submanifold sparse convolution. Stacking multi-layer regular sparse convolution to expand the receptive field would make the object contour more blurred (Figure 1(a)). The introduction of submanifold sparse convolution alleviates this problem. Focals Conv [4] adds a focus sparse convolution in Sparse CNNs to expand the receptive field of the object, whose output is suppressed in the foreground. But it is limited by the receptive field and cannot capture the long-range dependencies of voxels.

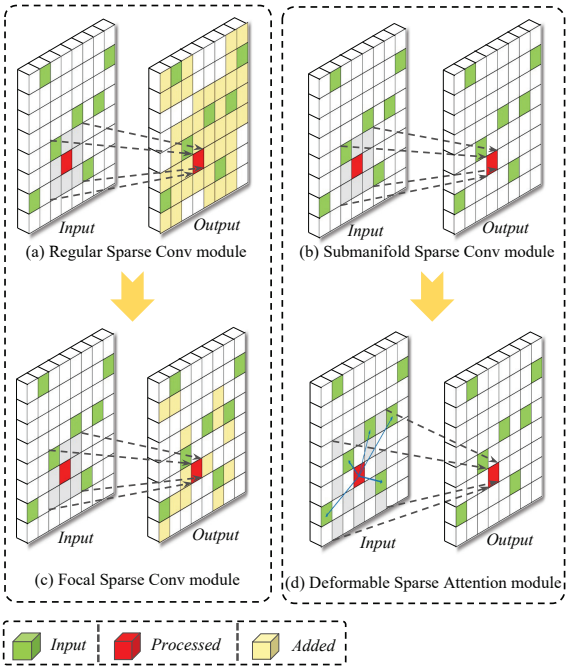

Fig. 1. Illustration of focal sparse convolution and deformable sparse attention. Regular sparse convolution seriously increases voxel feature density. Focus sparse convolution only expands information (yellow) on significant feature voxels, which preserves the sparsity of voxel features. Deformation sparse attention indicates that the query voxel (red) is sampled in an offset (green) manner based on the query kernel (gray). This 2D diagram can be easily extended to 3D forms.

Thanks to the global contextual information learning ability of the transformer, its variants are widely used in segmentation and detection tasks [5, 6]. VoTr [7] takes the non-empty voxels as input to transformer for 3D object detection task, and each query voxel performs the attention operation within a limited region. Inspired by [8], SST [9] divides the voxel

This work was supported in part by the National Key R&D Program of China under No.2021ZD0113503, Shanghai Municipal Science and Technology Major Project under No.2021SHZDZX0103. † Corresponding author.

feature map into groups and performs global information interaction between different groups through shift window operation. Although self-attention can learn long-range feature information, it weaken the local feature information. Up to now, the fusion of convolution and transformer is the most direct improvement method. Conformer [10], CvT [11], and D-DETR [12] combined two network paradigms and made significant improvements on 2D image tasks compared with the pure convolution method.

In addition, the importance of each non-empty voxel in the space for the 3D object detection task is different due to the sparsity of the point cloud. In the 2D image domain, the deformable methods [12, 13, 14] break the fixed geometric structure of the convolution and transformation paradigms, making the sampling focus on the object of our interest position, their various variant improve the modeling ability of geometric deformation objects. D-DETR [12] allows the query pixel to only interacts with offset sampling pixels. Centerformer [14] uses Sparse CNNs to convert 3D voxels into a 2D bird's-eye view (BEV) feature map, then follows [12] to learn features on the BEV map. These methods operate on 2D feature maps and cannot be directly applied to 3D sparse voxels.

To address the above problems, we propose a novel deformable sparse transformer (DST) augmented focal sparse convolution (FSC) 3D object detection backbone, termed D-Conformer. D-Conformer can replace Sparse CNNs module in most 3D detectors. We fuse local feature of the convolution and global feature of the transformer. FSC module (Figure 1(c)) replaces regular sparse convolution module in Sparse CNNs and enriches feature information only in the foreground. DST module (Figure 1(d)) replaces submanifold sparse convolution module. In deformable sparse attention of DST module, each voxel interacts only with non-empty voxels obtained by offset sampling, focusing on influential positions in sparse voxels. Our contributions can be summarized as follows:

(1) We propose a convolution and transformer fusion backbone for 3D object detection. The fusion backbone captures important position features from sparse voxels more resultful through local and long-range operations.

(2) We propose a deformable learning way for sparse voxel convolution and transform, which learns valuable voxels with position-wise in sparse space.

(3) Our D-Conformer outperforms previous state-of-the-art methods with a large margin, *i.e*, increasing 2.25% mAP in the hard car class on KITTI test set.

2. OUR APPROACH

2.1. Overall Architecture

We build up our focal sparse convolution and deformable sparse transformer fusion backbone, as shown in Figure 2. D-Conformer consists of a series of FSC and DST modules. D-Conformer uses FSC module (Figure 1(c)) to perform down-sampling on the voxel feature map, replacing regular sparse convolution (Figure 1a). Then two DST modules perform long-range information interaction on fine-grained local detail results aggregated by FSC. DST module (Figure 1(d)) have the same submanifold properties as submanifold sparse convolution (Figure 1(b)). Like Sparse CNNs [1], FSC module stacks three times to achieve multi-scale feature extraction.

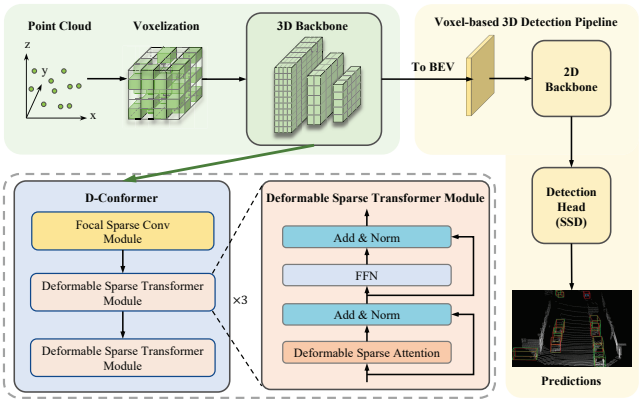

Fig. 2. Architecture overview for deformable sparse transformer augmented convolution (D-Conformer). D-Conformer is a 3D backbone that combines convolution and transformer, which can be applied in most voxel-based 3D detectors. It contains a series of focal sparse convolution and deformable sparse transformer modules.

2.2. Focal Sparse Convolution Module

Inspired by [4], we limit the output position of regular sparse convolution to more valuable foreground features (*i.e.*, making the foreground voxel dense while the background feature density remains unchanged mainly), as shown in Figure 1(c). The output of FSC module is obtained by pruning the cube importance feature map according to the threshold τ. The cube importance map I^p is predicted by an additional submanifold sparse convolution kernel and sigmoid function.

The output P_{out} is defined as a combination of all important positions, their extended regions and other unimportant positions, expressed as follows:

$$P_{\text{out}} = \left(\bigcup_{p \in P_{\text{im}}} P\left(p, K^d_{\text{im}}(p)\right) \right) \cup P_{\text{in/im}}, \qquad (1)$$

where function $\bigcup$ takes the union of all important positions and their extended area. P_{in} is the input, P_{im} is an important position set and a subset of P_{in}, $K^d_{\text{im}}(p)$ is the extended regions of P_{im}. P_{im} is calculated from:

$$P_{\text{im}} = \{p \mid I^p_0 \geq \tau, p \in P_{\text{in}}\}, 1 \geq \tau \geq 0, \qquad (2)$$

where I_0^p is the center of the cubic importance map at position p. The extended regions of P_{im} can be expressed as follows:

$$K_{\text{im}}^d(p) = \left\{ k \mid p + k \in P_{\text{in}}, I_k^p \geq \tau, k \in K^d \right\}. \quad (3)$$

where K^d is the 3^3 space centered at position p.

Dense 3D feature maps would bring unacceptable computation to the subsequent transformer module. The features of the remaining unimportant positions maintain the spatial geometric structure and are output by the submanifold sparse convolution operation.

We apply focal loss [15] as an objective loss function to supervise the importance prediction. Objective targets are the voxels inside 3D ground-truth boxes. The loss weight is set to 1.

2.3. Deformable Sparse Transformer Module

Since the receptive field of submanifold sparse convolution [1] is limited, we introduce a transformer to capture feature information between long-range voxels. In addition, the voxels in the 3D space are sparse and large, and the contribution of voxels to object detection is also different [4]. Therefore, we develop the deformable sparse attention based on deformable attention [12], which query voxel only interacts with voxels that are offset-sampled (Figure 1(d)). The query region is centered on the query voxel and sparsely extended outward, named query kernel. We use a linear projection layer to learn offset position $\Delta \boldsymbol{p}$. Like submanifold sparse convolution (Figure 1(b)), deformable sparse attention maintains the original 3D structure while expanding the receptive field.

The proposed deformable sparse attention replaces the self-attention in vanilla transformer (Figure 2). Given input voxel feature map $\boldsymbol{x}$, the multi-head deformable sparse attention (MDSA) as follows:

$$\text{MDSA}(\boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left(\sum_{n=1}^{N} A_{mn} \cdot V_{mn} \right), \quad (4)$$

where M and N are the head number and the total number of offset sampling voxels, $\boldsymbol{W}_m$ is learnable weight, A_{mn} is attention weight, and V_{mn} is attention value. A_{mn} is expressed as follows:

$$A_{mn} = \sigma \left(\boldsymbol{W}_{mn} \boldsymbol{x}_q \right), \quad (5)$$

where σ is softmax function,$\boldsymbol{W}_{mn}$ is learnable weight, $\boldsymbol{x}_q$ is the feature of a query voxel q.

Attention value V_{mn} is obtained by linear projection of features at offset sampling positions, expressed as follows:

$$V_{mn} = \boldsymbol{W}'_{mn} \boldsymbol{x} \left(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mn} \right). \quad (6)$$

where $\boldsymbol{W}'_{mn}$ is learnable weight, $\boldsymbol{p}_q$ is a 3-d reference point position, $\Delta \boldsymbol{p}_{mn}$ is the offset. The position at $\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mn}$ through down rounding.

The 3D voxels are indexed by a list, and cannot be indexed by coordinates like an image. We use voxel coordinate query scheme proposed in VoTr [7] to query the offset sampled non-empty voxels.

3. EXPERIMENT

3.1. Dataset

The KITTI dataset [16] is a commonly used 3D detection dataset. There are 7481 training samples and 7518 test samples. The training samples are divided into the train split with 3712 samples and the val split with 3769 samples. Mean Average Precision (mAP) is usually used as the official evaluation metric. mAP is calculated with recall 40 positions (R40) on test set and 11 recall positions (R11) on val split for comparison with other detectors. The result of ablation studies is calculated with mAP (R11) on val split.

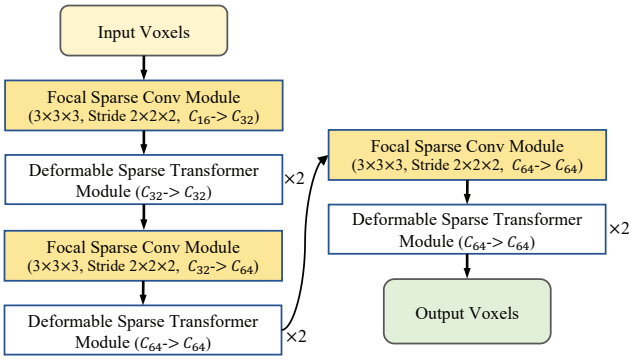

Fig. 3. Architecture of D-Conformer.

3.2. Implementation Details

We verify our module with the KITTI dataset in two popular point cloud 3D object detection frameworks, Second [1] and Voxel R-CNN [3]. We replace Sparse CNNs with D-Conformer as the new backbone. D-Conformer-SSD is a single-stage detector with Second as the framework. D-Conformer-TSD is a two-stage detector based on Voxel R-CNN. Inspired by [6], we first transform the original features of non-empty voxels through a linear projection layer into 16-channel initial features as input voxels. The detailed architecture of D-Conformer is shown in Figure 3. The feature dimension of the input voxel through the FSC module increased from 16 to 32, and remained unchanged in the DST module. In this paper, we set the offset sampling number of deformable sparse attention to 64 and follow [4] set the importance threshold τ to 0.5. Those networks are trained with batch size 2 on an NVIDIA RTX 3080 GPU. Due to sparse sampling, we set the dropout probability to 0.

Methods	backbone	Easy	Mod.	Hard
One-stage Methods				
SA-SSD [17]	C	88.75	79.79	74.16
VoTr-SSD [7]	T	86.73	78.25	72.99
Second [1]	C	84.65	75.96	68.71
D-Conformer-SSD (ours)	C+T	87.49	79.03	73.12
Two-stage Methods				
Part-A^2 [18]	C	87.81	78.49	73.51
PV-RCNN [2]	C	90.25	81.43	76.82
CT3D [19]	C	87.83	81.77	77.16
VoTr-TSD [7]	T	89.90	82.09	79.14
Focal Conv [4]	C	90.20	82.12	77.50
Voxel-RCNN [3]	C	90.90	81.62	77.06
D-Conformer-TSD (ours)	C+T	91.04	82.48	79.75

Table 1. Performance comparison on the KITTI test set with AP (R40) for car category. C+T denotes a detector using the fused Convolution and Transformer as the backbone. C denotes a detector using Convolution as the backbone.

Methods	Easy	Mod.	Hard
One-stage Methods			
SA-SSD [18]	90.15	79.91	78.78
VoTr-SSD [7]	87.86	78.27	76.93
Second [1]	88.61	78.62	77.22
D-Conformer-SSD (ours)	89.13	80.18	78.64
Two-stage Methods			
Part-A^2 [18]	89.47	79.47	78.54
PV-RCNN [2]	89.35	83.69	78.70
CT3D [19]	89.54	86.06	78.99
VoTr-TSD [7]	89.04	84.04	78.68
Focal Conv [4]	89.52	84.93	79.18
Voxel-RCNN [3]	89.41	84.52	78.93
D-Conformer-TSD (ours)	89.69	85.36	79.53

Table 2. Performance comparison on the KITTI val set with AP (R11) for car category.

3.3. Results on the KITTI dataset

In Table 1, Our method outperforms state-of-the-art methods Focal Conv [4] with a remarkable margin, *i.e.,* increasing 2.25% mAP on the hard car class of test set. The results of D-Conformer and VoTr [7] at the hard level illustrate the importance of long-range context information captured by the transformer for object detection. D-Conformer-SSD and D-Conformer-TSD outperform all the voxel-based detectors with a large margin, leading SECOND [1] and Voxel R-CNN [3] baselines by 3.07% and 0.86% on the moderate car class of test set. D-Conformer-TSD also surpasses transformer-based detector VoTr-TSD with 0.36% mAP absolute improvements. The above results indicate that fusing local features of convolution and long-range dependencies of transformer is helpful for 3D object detection. Table 2 shows the results on the KITTI val set. D-Conformer-SSD and D-Conformer-TSD outperform the baselines by a large margin 1.56% and 0.84% on the moderate car class. Overall, the results of test and val sets consistently show the effectiveness of D-Conformer.

Methods	F.S.C.	S.A.	D.S.A.	mAP
Voxel R-CNN	-	-	-	84.52
(a)	$\surd$	$\surd$		84.91
(b)	$\surd$		$\surd$	85.36

Table 3. Effects of proposed module on the moderate level car category of the KITTI val split with mAP (R11). F.S.C.: Focal Sparse Convolution. S.A.: Sparse Attention. D.S.T.: Deformable Sparse Attention.

Methods	kernel size / Sample Number	mAP
Voxel R-CNN	-	84.52
(a)	3 / 26	84.98
(b)	5 / 64	85.36

Table 4. Ablations on query kernel size in mAP (R11) on the moderate level car category of the KITTI val split.

3.4. Ablation Studies

Table 3 evaluates the impact of each proposed module on accuracy. Compared with the fusion of sparse transformer and FSC, the deformable sparse transformer fusion module can better extract sparse information and improve the result by 0.45% mAP absolute improvements. In particular, the ablation studies between regular sparse convolution and transformer are not carried out because the feature map obtained by regular sparse convolution is too dense.

Table 4 shows the impact of using different sizes of query kernel on D-Conformer. D-Conformer makes good improvements when the query kernel is 3^3. Note that we set the kernel 3^3 as the solid kernel because of its small receptive field. When query kernel is 5^3, the kernel is sparsely expanded, further improving the performance by 0.38% compared to kernel 3^3. The performance of D-Conformer increases with the enlargement of query kernel. But in general, the computational cost becomes unacceptable as the enlargement of the kernel.

4. CONCLUSION

This paper proposes a novel backbone network for 3D object detection based on the fusion of focal sparse convolution and deformable sparse transformer. D-Conformer performs feature extraction on sparse voxels more availably by learning important voxel radiation regions and advantageous non-empty voxel positions. Experimental results show that our proposed method outperforms most detectors based on pure convolution or transformer and verifies the effectiveness of D-Conformer on Second and Voxel R-CNN benchmark.

5. REFERENCES

[1] Yan Yan, Yuxing Mao, and Bo Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, pp. 3337, 2018.

[2] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10529–10538.

[3] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 1201–1209.

[4] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia, "Focal sparse convolutional networks for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5428–5437.

[5] Haihua Lu, Xuesong Chen, Guiying Zhang, Qiuhao Zhou, Yanbo Ma, and Yong Zhao, "Scanet: Spatial-channel attention network for 3d object detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1992–1996.

[6] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang, "3d object detection with pointformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7463–7472.

[7] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu, "Voxel transformer for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3164–3173.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[9] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang, "Embracing single stride 3d object detector with sparse transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8458–8468.

[10] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 367–376.

[11] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[13] Sudhir Yarram, Jialian Wu, Pan Ji, Yi Xu, and Junsong Yuan, "Deformable vistr: Spatio temporal deformable attention for video instance segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3303–3307.

[14] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh, "Centerformer: Center-based transformer for 3d object detection," *arXiv preprint arXiv:2209.05588*, 2022.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[16] Andreas Geiger, P Lenz, and R Urtasun, "Are we ready for autonomous driving," in *The KITTI Vision Benchmark Suite, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

[17] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11873–11882.

[18] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2647–2664, 2020.

[19] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao, "Improving 3d object detection with channel-wise transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2743–2752.