Vision Transformer Based Multi-class Lesion Detection in IVOCT

Zixuan Wang[1], Yifan Shao[2], Jingyi Sun[2], Zhili Huang[1], Su Wang[1(✉)], Qiyong Li[3], Jinsong Li[3], and Qian Yu[2(✉)]

[1] Sichuan University, Chengdu, China
hzl759156158@163.com
[2] Beihang University, Beijing, China
qianyu@buaa.edu.cn
[3] Sichuan Provincial People's Hospital, Chengdu, China

Abstract. Cardiovascular disease is a high-fatality illness. Intravascular Optical Coherence Tomography (IVOCT) technology can significantly assist in diagnosing and treating cardiovascular diseases. However, locating and classifying lesions from hundreds of IVOCT images is time-consuming and challenging, especially for junior physicians. An automatic lesion detection and classification model is desirable. To achieve this goal, in this work, we first collect an IVOCT dataset, including 2,988 images from 69 IVOCT data and 4,734 annotations of lesions spanning over three categories. Based on the newly-collected dataset, we propose a multi-class detection model based on Vision Transformer, called **G-Swin Transformer**. The essential part of our model is grid attention which is used to model relations among consecutive IVOCT images. Through extensive experiments, we show that the proposed G-Swin Transformer can effectively localize different types of lesions in IVOCT images, significantly outperforming baseline methods in all evaluation metrics. Our code is available via this link. https://github.com/Shao1Fan/G-Swin-Transformer

Keywords: IVOCT · Object Detection · Vision Transformer

1 Introduction

Despite the rapid development of new detection and treatment methods, the prevalence of cardiovascular disease continues to increase [1]. It is still reported to be the most prevalent and deadly disease worldwide, with more than 1 million people diagnosed with acute coronary syndrome (ACS) in the U.S. in 2016. The

Z. Wang and Y. Shao—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_32.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 327–336, 2023.
https://doi.org/10.1007/978-3-031-43987-2_32

average cost of hospital discharge for ACS patients is as high as $63,578 [2], which significantly increasing the financial burden on society and patients.

Optical coherence tomography (OCT) [3] is a new biomedical imaging technique born in the 19901990ss. Intravascular optical coherence tomography (IVOCT) [4] has a higher resolution compared with other imaging modalities in the vasculature and is considered to be the best imaging tool for plaque rupture, plaque erosion, and calcified nodules [5]. Therefore, most existing work on IVOCT images focuses on identifying vulnerable plaques in the vasculature [6–9], while neglecting other characteristic manifestations of atherosclerotic plaques in IVOCT images, such as macrophage infiltration and thrombus formation. These lesions are closely related to the development of plaque changes [10]. Studies have shown that atherosclerosis is an inflammatory disease dominated by macrophages and T lymphocytes, that a high density of macrophages usually represents a higher risk, and that thrombosis due to plaque rupture is a common cause of acute myocardial infarction [11,12]. In addition, some spontaneous coronary artery dissection (SCAD) can be detected in IVOCT images. The presence of the dissection predisposes to coronary occlusion, rupture, and even death [13,14]. These lesions are inextricably linked to ACS. All three types of features observed through IVOCT images are valuable for clinical treatment, as shown in Fig. 1. These lesions are inextricably linked to ACS and should be considered in clinical management.

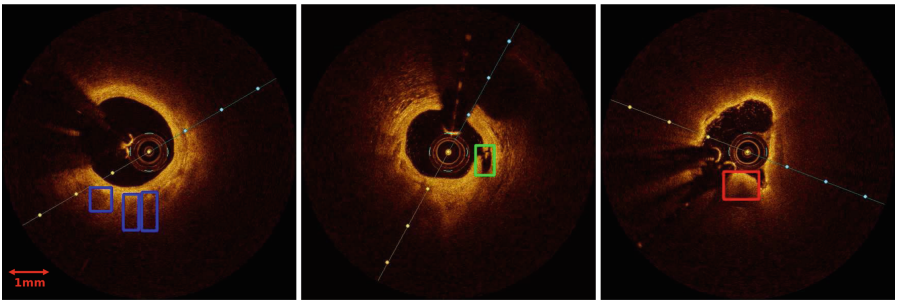

Fig. 1. Example images and annotations of our dataset. Each IVOCT data is converted to PNG images for annotation. The blue/green/red boxes represent bounding box of macrophages, cavities/dissections, thrombi, respectively. (Color figure online)

Achieving multi-class lesion detection in IVOCT images faces two challenges: 1) There is no public IVOCT dataset specifically designed for multi-class lesion detection. Most IVOCT datasets only focus on a single lesion, and research on the specific types of lesions in the cardiovascular system is still in its early stage. 2) It is difficult to distinguish between different lesions, even for senior radiologists. This is because these lesions vary in size and appearance within the same class, and some of them do not have regular form, as shown in Fig. 1. In clinical diagnosis, radiologists usually combine different pathological manifestations,

lesion size, and the continuous range before and after in the IVOCT image to design accurate treatment strategies for patients. Unfortunately, most existing works ignore such information and do not consider the continuity of lesions in the 3D dimension. To address the above issues, we collaborated with the Cardiovascular Research Center of Sichuan Provincial People's Hospital to collect an IVOCT dataset and introduce a novel detection model that leverages the information from consecutive IVOCT images.

Overall, the contribution of this work can be summarized as follows: 1) We propose a new IVOCT dataset that is the first multi-class IVOCT dataset with bounding box annotations for macrophages, cavities/dissections, and thrombi. 2) We design a multi-class lesion detection model with a novel self-attention module that exploits the relationship between adjacent frames in IVOCT, resulting in improved performance. 3) We explore different data augmentation strategies for this task. 4) Through extensive experiments, we demonstrate the effectiveness of our proposed model.

2 Dataset

We collected and annotated a new IVOCT dataset consisting of 2,988 IVOCT images, including 2,811 macrophages, 812 cavities and dissections, and 1,111 thrombi. The collected data from 69 patients are divided into training/validation/test sets in a 55:7:7 ratio, respectively. Each split contains 2359/290/339 IVOCT frames. In this section, we will describe the data collection and annotation process in detail.

2.1 Data Collection

We collaborated with the Cardiovascular and Cerebrovascular Research Center of Sichuan Provincial People's Hospital, which provided us with IVOCT data collected between 2019 and 2022. The data include OCT examinations of primary patients and post-coronary stenting scenarios. Since DICOM is the most widely-used data format in medical image analysis, the collecting procedure was exported to DICOM, and the patient's name and other private information contained in DICOM were desensitized at the same time. Finally, the 69 DICOM format data were converted into PNG images with a size of 575×575 pixels. It is worth noting that the conversion from DICOM to PNG did not involve any downsampling operations to preserve as much information as possible.

2.2 Data Annotation

In order to label the lesions as accurately as possible, we designed a two-step annotation procedure. The first round was annotated by two expert physicians using the one-stop medical image labeling software *Pair*. Annotations of the two physicians may be different. Therefore, we asked them to discuss and reach agreement on each annotation. Next, the annotated data was sent to senior doctors

to review. The review starts with one physician handling the labeling, including labeling error correction, labeling range modification, and adding missing labels. After that, another physician would continue to check and review the previous round's results to complete the final labeling. Through the above two steps, 2,988 IVOCT images with 4,734 valid annotations are collected.

3 Methodology

Recently, object detection models based on Vision Transformers have achieved state-of-the-art (SOTA) results on various object detection datasets, such as the MS-COCO dataset. Among them, the Swin Transformer [19] model is one of the best-performing models. Swin Transformer uses a self-attention mechanism within local windows to ensure computational efficiency. Moreover, its sliding window mechanism allows for global modeling by enabling self-attention computation between adjacent windows. Its hierarchical structure allows flexible modeling of information at different scales and is suitable for various downstream tasks, such as object detection.

3.1 G-Swin Transformer

In traditional object detection datasets such as the MS-COCO dataset, the images are typically isolated from each other without any correlation. However, in our proposed IVOCT dataset, each IVOCT scan contains around 370 frames with a strong inter-frame correlation. Specifically, for example, if a macrophage lesion is detected at the $[x, y, w, h]$ position in frame F_i of a certain IVOCT scan, it is highly likely that there is also a macrophage lesion near the $[x, y, w, h]$ position in frame F_{i-1} or F_{i+1}, due to the imaging and pathogenesis principles of IVOCT and ACS. Doctors also rely on the adjacent frames for diagnosis rather than a single frame when interpreting IVOCT scans. But, the design of the Swin-Transformer did not consider the utilization of inter-frame information. Though global modeling is enabled by using the sliding window mechanism. In the temporal dimension, it still has a locality because the model did not see adjacent frames.

Based on the Swin Transformer, we propose a backbone called G-Swin Transformer. Our proposed G-Swin Transformer is used as the basic module of the encoder in the full model, which is developed based on Faster R-CNN. The overall structure of the model is shown in Fig. 2. The model input consists of k 3-channel RGB images, and the input dimension is $[k * B, 3, H, W]$, where k indicates the number of frames that used in an iteration. After passing through Patch Partition and Linear Embedding layers, k feature maps belonging to frame F_0, F_1, ... F_{k-1}, respectively, are obtained, each with a size of H/4 * W/4 * C. These feature maps are then input to the G-Swin Transformer, where they go through 4 layers and a total of 12 Transformer blocks. Between each layer, a patch merging layer is used to reduce resolution, and model features of different dimensions. The output feature maps at different scales are then passed to a

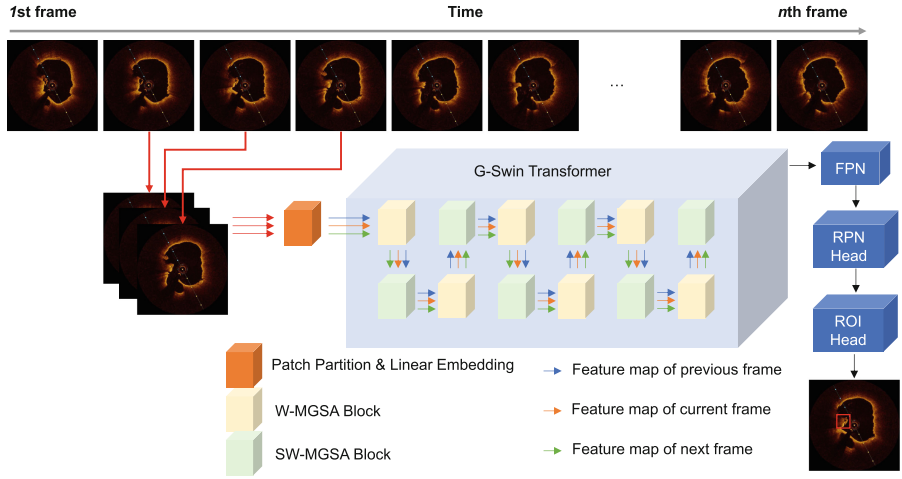

Fig. 2. The overall model structure. The proposed G-Swin Transformer is used as backbone network. The detection head follows Faster-RCNN's head. *W-MGSA* and *SW* refer to Window-Multihead Grid Self Attention and Shifted Window, respectively.

feature pyramid network (FPN) for fusion of features at different resolutions. The RPN Head is then applied to obtain candidate boxes, and finally, the ROI Head is used for classification and refinement of candidate boxes to obtain class and bbox (bounding box) predictions. The inter-frame feature fusing is happend in the attention block, introduced in the next subsection.

3.2 Grid Attention

To better utilize information from previous and future frames and perform feature fusion, we propose a self-attention calculation mode called "Grid Attention". The structure shown in Fig. 3 is an application of Grid Attention. The input of the block is 3 feature maps respectively from frames 0, 1, and 2. (Here we use $k = 3$.) Before entering the W-MSA module for multi-head self-attention calculation, the feature maps from different frames are fused together.

Based on the feature map of the key frame (orange color), the feature maps of the previous (blue) and next (green) frames first do a dimensional reduction from $[H, W, C]$ to $[H, W, C/2]$. Then they are down-sampled and a grid-like feature map are reserved. The grid-like feature map are then added to key-frame feature map, and the fusion progress finishes. In the W-MSA module, the self-attention within the local window and that between adjacent local windows are calculated, and the inter-frame information is fully used. The local window of key-frame has contained information from other frames, and self-attention calculation happens in inter-frames. The frame-level feature modeling can thus be achieved, simulating the way that doctors view IVOCT by combining information from previous and next frames.

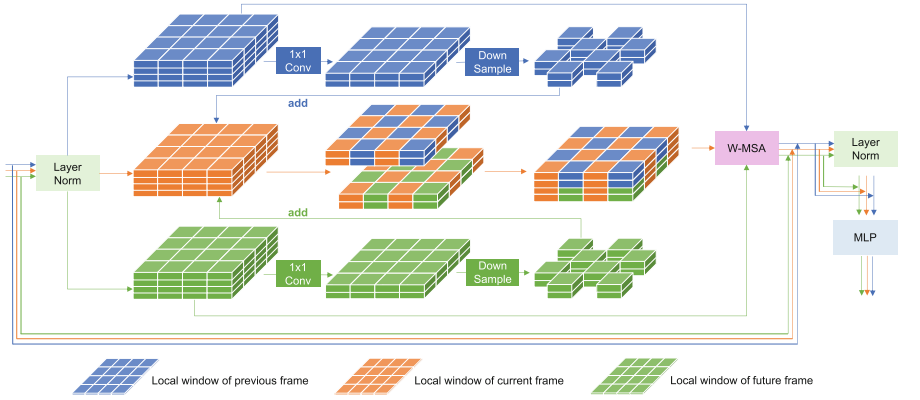

Fig. 3. Illustration of the proposed Grid Attention. The blue/orange/green feature map belongs to a local window of the previous/current/next frame. After the dimensional reduction and downsampling operation, the feature maps of previous/next frame is added to the current frame's feature map. (Color figure online)

During feature fusion with Grid Attention, the feature maps from different frames are fused together in a grid-like pattern (as shown in the figure). The purpose of this is to ensure that when dividing windows, half of the grid cells within a window come from the current frame, and the other half come from other frames. If the number of channels in the feature map is C, and the number of frames being fused is 3 (current frame + previous frame + next frame), then the first C/2 channels will be fused between the current frame and the previous frame, and the last C/2 channels will be fused between the current frame and the next frame. Therefore, the final feature map consists of 1/4 of the previous frame, 1/2 of the current frame, and 1/4 of the next frame. The impact of the current frame on the new feature map remains the largest, as the current frame is the most critical frame.

4 Experiments

Baseline Methods and Evaluation Metrics. The baseline is based on a PyTorch implementation of the open-source object detection toolbox MMDetection. We compare our proposed approach with Swin Transformer and four CNN-based network models including Faster-RCNN [15], YOLOv3 [16], YOLOv5 [17], Retinanet [18]. All the baseline model is pre-trained on the ImageNet dataset.

To ensure objective comparison, all experiments were conducted in the MMdetection framework. The metric we used is the AP/AR for each lesion and the mAP, based on the COCO metric and the COCO API (the default evaluation method in the MMdetection framework). We trained the model for 60 epochs with an AdamW optimizer following Swin Transformer. The learning

Table 1. Comparison of our proposed method and baseline methods.

	AP_{50}			mAP	$Recall_{50}$		
Method	Macrophage	cavities/ dissection	thrombus		Macrophage	cavities/ dissection	thrombus
Faster-RCNN	27.34	44.32	31.86	34.51	74.65	76.04	70.60
YOLOv3	20.25	35.42	33.17	29.61	67.37	75.52	61.82
YOLOv5	27.34	40.63	46.93	38.30	79.31	82.67	83.86
RetinaNet	25.86	38.93	30.17	31.65	84.48	88.54	73.03
Swin-Transformer	27.91	44.94	48.87	40.57	89.11	89.06	92.85
G-Swin-Transformer (Ours)	**30.55**	**52.25**	**51.92**	**44.91**	**91.49**	**89.58**	**95.45**

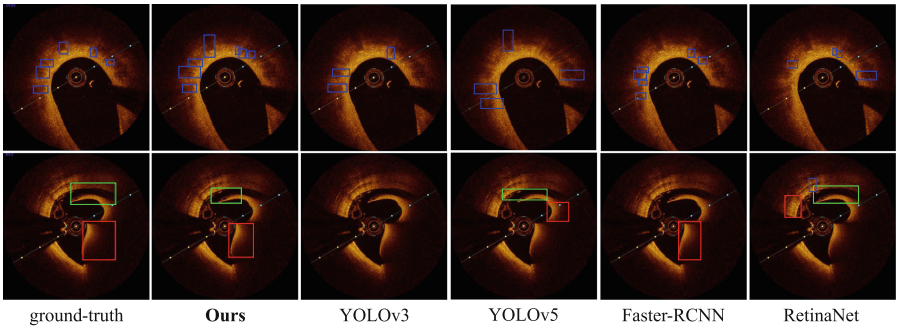

Fig. 4. Visualization results. From left to right are ground-truth, results of our model, Faster-RCNN, YOLOv3, YOLOv5 and RetinaNet. Our model achieves better results.

rate and weight decay is set to be 1e-4 and 1e-2, respectively. The batch size is set to be 2.

Quantitative and Qualitative Results. All experiments are conducted on our newly-collected dataset. Each model is trained on the training set, selected based on the performance of the validation set, and the reported results are obtained on the test set. Table 1 shows the comparison between the baseline methods and the G-Swin Transformer method. Methods based on Swin Transformer outperformed the four baseline methods in terms of precision and recall, and our proposed G-Swin Transformer outperforms the baseline method Swin Transformer by 2.15% in mAP.

Figure 4 compares some results of our method and baselines. The first row is the detection of the macrophage. Our method's prediction is the most closed to the ground truth. The second row is the detection of cavities/dissections and thrombi. Only our method gets the right prediction. The YOLOv3, YOLOv5, Faster-RCNN and RetinaNet model failed to detect all lesions, while RetinaNet model even produced some false positive lesions.

Table 2. Results of using different data augmentation strategies.

Augmentation Strategy					Metric	
Random Resize	Random Crop	Random Flip	Random Brightness	Random Contrast	mAP	AR
✗	✗	✔	✔	✔	33.13	79.12
✗	✔	✔	✔	✔	35.64	82.31
✔	✗	✔	✔	✔	38.52	85.81
✔	✔	✔	✗	✗	41.31	85.69
✔	✔	✔	✗	✔	41.92	85.31
✔	✔	✔	✔	✗	42.39	86.17
✔	✔	✗	✔	✔	43.34	91.41
✔	✔	✔	✔	✔	**44.91**	**92.18**

Table 3. Effect of different hyper-parameters in Grid Attention.

Fusion Layers	Fusion Strategy	mAP	AR
5	replace	41.69	89.28
5	add	40.31	90.52
3	replace	44.39	88.86
3	**add**	**44.91**	**92.18**

Table 4. Comparison of different fusion strategies.

Fusion methods	mAP	AR
No fusion	40.57	90.34
2.5D	38.25	78.44
Weighted sum	40.64	83.54
Ours	**44.91**	**92.18**

Effect of Different Data Augmentation Methods. We compared the impact of different data augmentation strategies in our task. As shown in Table 2, *Random Resize* and *Random Crop* had a significant impact on performance improvement. *Resize* had the greatest impact on the model's performance because different-sized OCT images were generated after data augmentation, and the lesions were also enlarged or reduced proportionally. Since the sizes of lesions in different images are usually different, different-sized lesions produced through data augmentation are advantageous for the model to utilize multi-scale features for learning.

Effect of Different Hyper-parameters. Table 3 shows the impact of hyper-parameters on the performance of the G-Swin Transformer model. The best mAP was achieved when using a 3-layer image input. Using the upper and lower 5 layers of image input not only increased the training/inference time, but also

may not provide more valuable information since frame 0 and frame 4 are too far away from the key frame. The fusion strategy indicates how the feature map from other frames are combined with the key-frame feature map. We can find add them up gets better result then simply replacement. We think this is because by this way, the 1×1 convolutional layer can learn a residual weights, keeps more detail of the key-frame.

Effect of Fusion Methods. In addition to Grid Attention, there are other methods of feature fusion. The first method is like 2.5D convolution, in which multiple frames of images are mapped into 96-dimensional feature maps directly through convolution in the Linear Embedding layer. This method is the simplest, but since the features are fused only once at the initial stage of the network, the use of adjacent frame features is very limited. The second method is to weight and sum the feature maps of different frames before each Attention Block, giving higher weight to the current frame and lower weight to the reference frames. Table 4 shows the impact of other feature fusion methods on performance. Our method gets better mAP and AR.

5 Conclusion

In this work, we have presented the first multi-class lesion detection dataset of IVOCT scans. We have also proposed a Vision Transformer-based model, called G-Swin Transformer, which uses adjacent frames as input and leverages the temporary dimensional information inherent in IVOCT data. Our method outperforms traditional detection models in terms of accuracy. Clinical evaluation shows that our model's predictions provide significant value in assisting the diagnosis of acute coronary syndrome (ACS).

Acknowledgement. This work is supported by the National Key Research and Development Project of China (No. 2022ZD0117801).

References

1. Murphy, S., Xu, J., Kochanek, K., Arias, E., Tejada-Vera, B.: Deaths: final data for 2018 (2021)
2. Virani, S., et al.: Heart disease and stroke statistics-2021 update: a report from the American heart association. Circulation. **143**, e254–e743 (2021)
3. Huang, D., et al.: Optical coherence tomography. Science **254**, 1178–1181 (1991)
4. Bezerra, H., Costa, M., Guagliumi, G., Rollins, A., Simon, D.: Intracoronary optical coherence tomography: a comprehensive review: clinical and research applications. JACC: Cardiovas. Interv. **2**, 1035–1046 (2009)
5. Jia, H., et al.: In vivo diagnosis of plaque erosion and calcified nodule in patients with acute coronary syndrome by intravascular optical coherence tomography. J. Am. Coll. Cardiol. **62**, 1748–1758 (2013)
6. Li, C., et al.: Comprehensive assessment of coronary calcification in intravascular OCT using a spatial-temporal encoder-decoder network. IEEE Trans. Med. Imaging **41**, 857–868 (2021)

7. Liu, X., Du, J., Yang, J., Xiong, P., Liu, J., Lin, F.: Coronary artery fibrous plaque detection based on multi-scale convolutional neural networks. J. Signal Process. Syst. **92**, 325–333 (2020)
8. Gessert, N., et al.: Automatic plaque detection in IVOCT pullbacks using convolutional neural networks. IEEE Trans. Med. Imaging **38**, 426–434 (2018)
9. Cao, X., Zheng, J., Liu, Z., Jiang, P., Gao, D., Ma, R.: Improved U-net for plaque segmentation of intracoronary optical coherence tomography images. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021. LNCS, vol. 12893, pp. 598–609. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86365-4_48
10. Regar, E., Ligthart, J., Bruining, N., Soest, G.: The diagnostic value of intracoronary optical coherence tomography. Herz: Kardiovaskulaere Erkraenkungen **36**, 417–429 (2011)
11. Kubo, T., Xu, C., Wang, Z., Ditzhuijzen, N., Bezerra, H.: Plaque and thrombus evaluation by optical coherence tomography. Int. J. Cardiovasc. Imaging **27**, 289–298 (2011)
12. Falk, E., Nakano, M., Bentzon, J., Finn, A., Virmani, R.: Update on acute coronary syndromes: the pathologists' view. Eur. Heart J. **34**, 719–728 (2013)
13. Saw, J.: Spontaneous coronary artery dissection. Can. J. Cardiol. **29**, 1027–1033 (2013)
14. Pepe, A., et al.: Detection, segmentation, simulation and visualization of aortic dissections: a review. Med. Image Anal. **65**, 101773 (2020)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances In Neural Information Processing Systems, vol. 28 (2015)
16. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. ArXiv Preprint ArXiv:1804.02767 (2018)
17. Jocher, G.: YOLOv5 by ultralytics (2020). https://github.com/ultralytics/yolov5
18. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference On Computer Vision, pp. 2980–2988 (2017)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference On Computer Vision, pp. 10012–10022 (2021)