ESGN: Efficient Stereo Geometry Network for Fast 3D Object Detection

Aqi Gao, Yanwei Pang, *Senior Member, IEEE,* Jing Nie, Zhuang Shao, Jiale Cao, Yishun Guo, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Fast stereo based 3D object detectors have made great progress recently. However, they suffer from the inferior accuracy. We argue that the main reason is due to the poor geometry-aware feature representation in 3D space. To solve this problem, we propose an efficient stereo geometry network (ESGN). The key in our ESGN is an efficient geometry-aware feature generation (EGFG) module. Our EGFG module first uses a stereo correlation and reprojection module to construct multi-scale stereo volumes in camera frustum space, second employs a multi-scale bird's eye view (BEV) projection and fusion module to generate multiple geometry-aware features. In these two steps, we adopt deep multi-scale information fusion for discriminative geometry-aware feature generation, without any complex aggregation networks. In addition, we introduce a deep geometry-aware feature distillation scheme to guide stereo feature learning with a LiDAR-based detector. The experiments are performed on the classical KITTI dataset. On KITTI test set, our ESGN outperforms the fast state-of-art-art detector YOLOStereo3D by 5.14% on mAP$_{3d}$ at 62ms. To the best of our knowledge, our ESGN achieves a best trade-off between accuracy and speed. We hope that our efficient stereo geometry network can provide more possible directions for fast 3D object detection.**

***Index Terms*—Autonomous driving, 3D detection, Stereo images, Computer vision.**

I. Introduction

3D object detection is an important but challenging computer vision task, which is essential for automatic driving. Though LiDAR-based 3D object detection approaches [33], [34], [39] have high accuracy, they suffer from the expensive hardware cost and low resolution. Compared with LiDAR-based 3D object detection approaches, stereo-based 3D object detection approaches [24], [30], [47] adopt the low-cost optical camera and can provide dense 3D information. The stereo-based 3D object detection approaches can be mainly divided into camera frustum space based methods [27], pseudo LiDAR based methods [20], [21], and voxel based methods [11], [15].

Currently, fast stereo 3D methods belong to camera frustum space based methods and pseudo LiDAR based methods. As shown in Fig. 1(I), camera frustum space based method YOLOStereo3D [27] first employs a stereo 3D/4D correlation module to generate stereo volume in camera frustum space, and second performs 3D object detection directly on stereo volume. Pseudo LiDAR based methods first generate point cloud with estimated depth and then employ a light-weighted LiDAR-based detector for 3D detection. However, camera frustum space based methods lack effective feature representation in 3D space, resulting in the issue of object distortion, while pseudo LiDAR based methods are sensitive to the precision of fast depth estimation network. Compared with these fast methods, voxel based methods are dominant in accuracy. Fig. 1(II) shows the pipeline of voxel based method DSGN [11]. It first uses a plane sweep and aggregation module to generate stereo volume in camera frustum space. After that, it applies a deep bird's eye view (BEV) projection and aggregation module to extract geometry-aware feature in 3D space, and performs 3D detection. In these two steps above, voxel based methods employ heavy 3D and 2D aggregation networks to extract discriminative features, which leads to a slow speed. Namely, these existing methods do not achieve a good trade-off between speed and accuracy. Therefore, it is important to design a effective and efficient stereo 3D detector.

To achieve this goal, we propose an efficient stereo geometry network (ESGN) in Fig. 1(III). The key is an efficient 3D geometry-aware feature generation module (EGFG), which consists of a stereo correlation and reprojection (SCR) module, and a multi-scale BEV projection and fusion (MPF) module. We first use the SCR module to generate multi-scale stereo volumes in camera frustum space. After that, we apply the MPF module to convert multi-scale stereo volumes into multiple geometry-aware features in 3D space. Finally, we perform 3D detection on one of geometry-aware features. In addition, we introduce a deep geometry-aware feature distillation scheme to guide feature learning, where a LiDAR-based detector is designed to provide deep supervisions in multiple levels. Compared to these existing stereo 3D methods, our proposed ESGN achieves an optimal trade-off between accuracy and speed. We hope that our proposed method can provide more possible directions for fast stereo 3D object detection, and promote stereo application to automatic driving and robot. Our contributions and metrics can be summarized as follows:

- We propose an efficient stereo geometry-aware feature

This work was supported by National Key Research and Development Program of China (No. 2018AAA0102800, 2018AAA0102802), National Natural Science Foundation of China (No. 61906131), Tianjin Science and Technology Program (No. 19ZXZNGX00050), Tianjin Natural Science Foundation (No. 21JCQNJC00420), and CAAI-Huawei MindSpore Open Fund.

A. Gao, Y. Pang, J. Nie, J. Cao and Y. Guo are with the Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (E-mail: {gaoaqi,pyw,jingnie,connor,guoyishun}@tju.edu.cn). (Corresponding author: Yanwei Pang and Jiale Cao)

Z. Shao is in Warwick Manufacturing Group at University of Warwick, Coventry, UK (E-mail: Zhuang.Shao@warwick.ac.uk).

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (E-mail: xuelong_li@outlook.com).

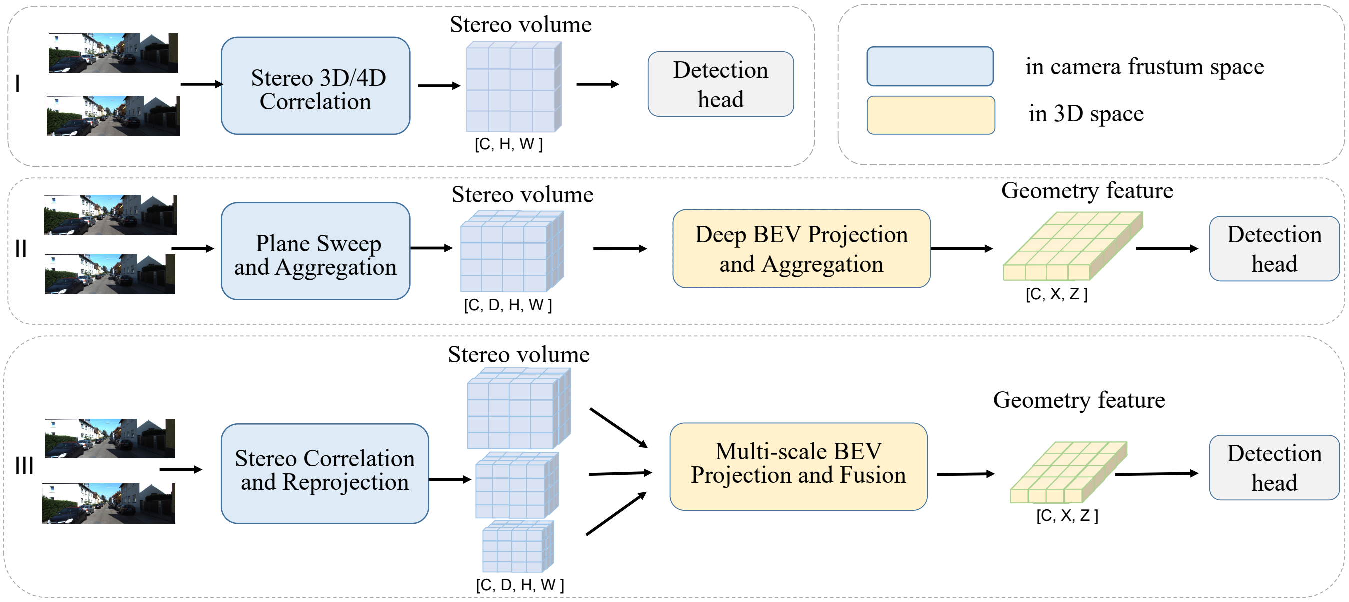


Fig. 1. Comparison of different stereo methods. (I) Camera frustum space based method YOLOStereo3D [27]. It uses a stereo 3D/4D correlation module to generate stereo volume in camera frustum space for 3D detection, which faces the issue of object distortion. (II) Voxel based method DSGN [11]. In the steps of stereo volume and geometry-aware feature generations, it adopts the heavy 3D and 2D aggregation networks, resulting in a slow speed. (III) Our efficient stereo geometry-aware network (ESGN). Compared to voxel based method DSGN, our ESGN adopts deep multi-scale information fusion, instead of heavy 3D and 2D aggregation networks, for geometry-aware feature generation.

network (ESGN) for fast 3D object detection. The key module is an efficient geometry-aware feature generation (EGFG) module. EGFG extracts discriminative geometry-aware features in 3D space by adopting deep multi-scale information fusion in both stereo volume and geometry-aware feature generations.

- We introduce a deep geometry-aware feature distillation (DGFD) scheme. DGFD uses a LiDAR-based detector to extract multi-level geometry-aware features and employs these features to guide stereo feature learning.
- We perform the experiments on the classical KITTI dataset [14]. On moderate test set, our ESGN achieves an AP_{3d} of 46.39% at a speed of $62ms$, obtaining an optimal trade-off between accuracy and speed. Compared to fast YOLOStereo3D [27], our ESGN provides an absolute gain of 5.14% at a comparable speed.

II. Related Work

Compared to 2D object detection [7], [13], [36], [56], 3D object detection [11], [24], [27] aims to classify and localize objects in 3D space, which is more challenging and useful for real applications. In this section, we first introduce stereo 3D object detection. After that, we review LiDAR-based 3D object detection, and feature representation and distillation.

A. Stereo 3D Object Detection

As mentioned earlier, stereo 3D object detection can be mainly divided into three classes: camera frustum space based methods, pseudo LiDAR based methods, voxel based methods. Camera frustum space represents a frustum from the perspective camera system, where a point coordinate is (u, v, d_{3d}), where u, v are the pixel coordinate in the image space and d_{3d} is depth coordinate. Therefore, camera frustum space based methods extract the features in image coordinate system for 3D object detection. Stereo RCNN [24] first predicts a rough 3D bounding box based on the combined RoI features from the left and right images, and second conducts a bundle adjustment optimization for final 3D bounding box prediction. IDA-3D [30] builds a cost volume from left and right RoI features to predict the depth of center point for 3D object detection.

Pseudo LiDAR based methods convert stereo 3D detection into LiDAR-based 3D detection. Pseudo-LiDAR [47] is one of the earliest pseudo LiDAR based methods. It first uses stereo matching methods (such as [2], [4], [5]) to get disparity map, and then transforms the disparity map into the point cloud data, and finally performs 3D point cloud detection. Pseudo-LiDAR++ [54] introduces a depth cost volume to generate depth map directly. OC-Stereo [31] and Disp RCNN [45] only consider the foreground regions of point cloud and achieve a better performance. ZoomNet [49] improves the disparity estimation by enlarging the target region. The work of [1] applies Pseudo LiDAR based method into road detection.

Voxel based methods extract voxel features in 3D space to detect objects. Compared to camera frustum space, 3D space can avoid object distortion. DSGN [11] converts stereo volume in camera frustum space into the volume in 3D space to better represent 3D object structure. LIGA [15] applies a LiDAR-based detector as a teacher model to guide stereo feature learning. With discriminative geometry-aware features, these methods are dominate in accuracy. However, due to heavy 3D convolutions and heavy 3D/2D aggregation networks, these methods are insufficient enough to be applied in practice.

Fast stereo 3D object detection approaches mainly belong

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3202810

to camera frustum space based and pseudo LiDAR based methods. Pseudo LiDAR based methods RT3D-GMP [21] and RT3DStereo [20] use a light-weighted depth estimation module to generate depth map. After that, they transform the depth map into the point cloud and use a light-weighted 3D point cloud detector to perform 3D detection. With these light-weighted modules, these methods have a high speed. Camera frustum space based method Stereo-Centernet [42] changes the anchor-based network in Stereo-RCNN [24] into a key-point based network for fast 3D detection. YOLOStereo3D [27] first adopts a stereo 3D/4D correlation module to generate stereo volume in camera frustum space and second performs 3D detection directly on stereo volume. YOLOStereo3D achieves best performance among these fast methods. However, due to the poor 3D geometry-aware feature representation, YOLOStereo3D lags far behind voxel based methods [11] in accuracy. Our proposed method aims to bring this gap by efficiently extracting discriminative geometry-aware feature and achieve a best trade-off between accuracy and speed.

B. LiDAR-based 3D Object Detection

Compared to stereo 3D detection, LiDAR-based 3D object detection has a higher accuracy, but suffers from expensive hardware cost and low resolution. LiDAR-based 3D object detection can be mainly divided into three classes: voxel based methods, point based methods, point-voxel based methods. Voxel based methods [22], [40], [53], [62] transform the irregular point clouds to volumetric representation in compact shape and extract the voxel features for 3D detection. Some backbones (*e.g.,* PointNet [33] and PointNet++ [34]) are proposed to directly extract the features from irregular point cloud data. Based on these backbones, point based methods [32], [39], [41], [48], [51] directly perform 3D detection. Voxel based methods are usually efficient but face the issue of information loss, while point based methods have a large receptive field but are inefficient. Point-voxel based methods [10], [16], [38], [52] aim to integrate the advantages of voxel based and point based methods. T3D [3] designs a transformer based vote refinement module to improve 3D detection.

C. Feature Representation and Distillation

Feature representation plays a key role in computer vision tasks. At first, handcrafted features are widely used in detection [6], [12], classification [29], and segmentation [8]. Recently, deep convolutional neural networks are proposed to extract deep features [17], [43], [59]. Compared to handcrafted features, deep features are more discriminative. With deep features, the computer vision tasks make a great progress in recent years, such as detection [24], [44], [57], segmentation [9], [55], and classification [58], [60]. These methods improve feature representation via better network or architecture design.

Knowledge distillation is first proposed for network compression [19], where a large and high-performance teacher network provides softened labels to supervise feature learning of a small student network. As a result, the student network can learn better features with a small number of network parameters. After that, some methods [18], [37] explore to make use of the knowledge from the intermediate layers of teacher network. Recently, knowledge distillation has been successfully applied to stereo 3D object detection [15], which demonstrates that it is effective to use LiDAR-based detector to guide stereo feature learning. We argue that single-level distillation in [15] can not provide deep supervision for stereo feature learning. To solve this issue, we propose a deep geometry-aware feature distillation scheme that provides deep multi-level supervisions on stereo feature learning.

III. Our Method

In this section, we introduce our efficient stereo geometry network (ESGN) for 3D object detection. Fig. 2(a) shows the overall architecture of our proposed ESGN. We first employ an efficient deep model ResNet-34 [17] to extract multi-scale paired feature maps ($\{F_l^i,F_r^i\}$, $i=1,2,3$) from stereo input images. Based on these paired feature maps, we design a novel efficient geometry-aware feature generation (EGFG) module to generate multiple geometry-aware features (F_{gf}^i, $i=1,2,3$) with deep multi-scale information fusion. Our EGFG module contains a stereo correlation and reprojection (SCR) module, and a multi-scale BEV preservation and fusion (MPF) module. Specifically, we first use a SCR module to generate stereo volume, which represents 3D geometry-aware feature in camera frustum space. To avoid the object distortion in camera frustum space, we then adopt a MPF module to transform stereo volume in camera frustum space into BEV feature in 3D world space, where the BEV feature contains 3D geometry-aware features in 3D space. With the joint SCR and MPF modules, our proposed method is able to perceive 3D geometry-aware features of objects. To further enhance geometry-aware feature representation, we introduce a deep geometry-aware feature distillation (DGFD) scheme, where a LiDAR-based detector is designed to extract multi-scale discriminative geometry-aware features (F_{lgf}^i, $i=1,2,3$) from point cloud data and then guide stereo geometry-aware feature learning. Finally, the geometry-aware feature F_{gf}^3 is fed to 3D prediction head for 3D object detection.

Here, we first introduce the key efficient geometry-aware feature generation (EGFG) module in Sec. III-A. After that, we describe deep geometry-aware feature distillation (DGFD) scheme in Sec. III-B. Finally, we describe the prediction heads, including 3D detection head and auxiliary heads, in Sec. III-C.

A. Efficient Geometry-Aware Feature Generation (EGFG)

The efficient geometry-aware feature generation (EGFG) module converts multi-scale paired feature maps ($\{F_l^i,F_r^i\}$, $i=1,2,3$) extracted from stereo images to multiple geometry-aware features (F_{gf}^i, $i=1,2,3$) with deep multi-level fusion. The EGFG module consists of two sequential modules: a SCR module and a MPF module. The SCR module first generates multi-scale stereo volumes in camera frustum space with simple correlation and repojection operations. Then, the MPF module converts multi-scale stereo volumes into 3D and BEV spaces to generate multiple geometry-aware features.

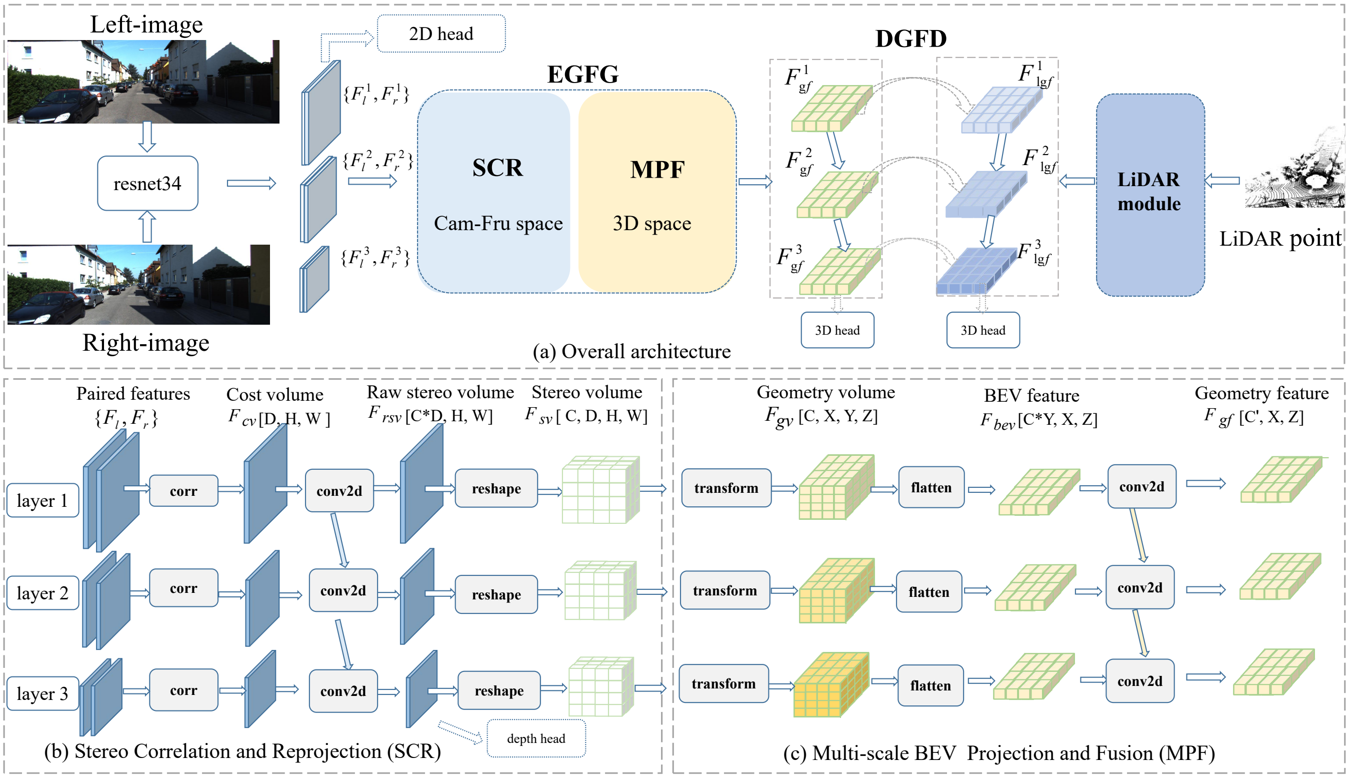

Left-image
resnet34
Right-image
2D head
$\{F_l^1, F_r^1\}$
$\{F_l^2, F_r^2\}$
$\{F_l^3, F_r^3\}$
EGFG
SCR
Cam-Fru space
MPF
3D space
DGFD
F_{gf}^1
F_{gf}^2
F_{gf}^3
F_{lgf}^1
F_{lgf}^2
F_{lgf}^3
3D head
3D head
LiDAR module
LiDAR point
(a) Overall architecture
Paired features $\{F_l, F_r\}$
Cost volume F_{cv}[D, H, W]
Raw stereo volume F_{rsv} [C*D, H, W]
Stereo volume F_{sv} [C, D, H, W]
layer 1
layer 2
layer 3
corr
conv2d
reshape
depth head
(b) Stereo Correlation and Reprojection (SCR)
Geometry volume F_{gv} [C, X, Y, Z]
BEV feature F_{bev}[C*Y, X, Z]
Geometry feature F_{gf} [C', X, Z]
transform
flatten
conv2d
(c) Multi-scale BEV Projection and Fusion (MPF)

Fig. 2. (a) Overall architecture of our ESGN. Given a paired of input images, we adopt the efficient ResNet-34 [17] to extract multi-scale paired feature maps ($\{F_l^1, F_r^1\}, i = 1, 2, 3$). Then, we employ our proposed efficient geometry-ware feature generation (EGFG) module to generate multiple geometry-ware features ($F_{gf}^i, i = 1, 2, 3$). In addition, we introduce a LiDAR-based detector for deep geometry-aware feature distillation (DGFD). (b) Stereo correlation and reprojection (SCR) module converts multi-scale paired feature maps to multiple stereo volumes in camera frustum space. (c) Multi-scale BEV projection and fusion (MPF) module converts camera frustum space to 3D space for geometry-aware feature generation.

1) Stereo Correlation and Reprojection (SCR): As shown in Fig. 2(b), the SCR module takes multi-scale paired features ($\{F_l^i,F_r^i\}$, $i = 1, 2, 3$) as inputs. With these paired features, the SCR module first extracts multi-scale cost volumes (F_{cv}^i, $i = 1, 2, 3$) with stereo correlation operation, and second generates multi-scale stereo volumes (F_{sv}^i, $i = 1, 2, 3$) in camera frustum space with reprojection.

For a paired feature map $\{F_l^i,F_r^i\}$, the cost volume (F_{cv}^i) is first generated with stereo correlation as follows.

$$F_{cv}^i(d, h, w) = \frac{1}{C}\sum_{c=1}^{C} F_l^i(c, h, w - d) * F_r^i(c, h, w + d), \tag{1}$$

where d, h, w are the indexes of disparity, height, and width dimensions, and C is the number of feature channels. The disparity index d contains depth information, which is inversely proportional to the depth. $w-d$ represents right shifting d pixels for left feature map, while $w + d$ represents left shifting d pixels for right feature map. Eq. 1 is a classical cost volume generation strategy [28], [46] that calculates cost volume features by greed search in disparity space. For the disparity d, the feature in left image is $F_l^i(c, h, w - d)$, and the feature in right image is $F_l^i(c, h, w+d)$. We calculate cost volume features by cross-channel feature correlation. The cost volume features have strong response at index d if a 3D point has a disparity d on stereo images. If $w - d$ or $w + d$ are out of range of F_l^i and F_r^i, we set the feature as zero.

After that, we convert multi-scale cost volumes (F_{cv}^i, $i = 1, 2, 3$) to multi-scale stereo volumes ($F_{sv}^i, i = 1, 2, 3$) with our repojection operation, including multi-scale cost volume fusion and dimension reshaping. Assuming that the size of F_{cv}^1 is $D \times H \times W$, we first use a 2D convolution to generate raw stereo volume $F_{rsv}^1 \in \mathbb{R}^{(C*D)\times H\times W}$ and second reshape F_{rsv}^1 to stereo volume $F_{sv}^1 \in \mathbb{R}^{C\times D\times H\times W}$.The 2D convolution is able to generate stereo volume based on depth information existing in cost volume. At the same time, we downsample raw stereo volume F_{rsv}^1 twice time and concatenate it with F_{cv}^2, which are fed to a 2D convolution to generate raw stereo volume $F_{rsv}^2 \in \mathbb{R}^{(C*D)\times H/2\times W/2}$. The F_{rsv}^2 is reshaped to generate stereo volume $F_{sv}^2 \in \mathbb{R}^{C\times D\times H/2\times W/2}$. Similarly, we generate stereo volume F_{sv}^3 with the inputs of F_{rsv}^2 and F_{cv}^3. The above steps of our reprojection operation can be summarized as the following equations.

$$\begin{cases} F_{sv}^1 = f_{re}(F_{rsv}^1), F_{rsv}^1 = f_{conv}(F_{cv}^1), \\ F_{sv}^2 = f_{re}(F_{rsv}^2), F_{rsv}^2 = f_{conv}(f_{cat}(f_{avg}(F_{rsv}^1), F_{cv}^2)), \\ F_{sv}^3 = f_{re}(F_{rsv}^3), F_{rsv}^3 = f_{conv}(f_{cat}(f_{avg}(F_{rsv}^2), F_{cv}^3)), \end{cases} \tag{2}$$

where f_{re} represents dimension reshaping, f_{conv} represents 2D convolution, f_{cat} represents channel concatenation, f_{avg} represents downsampling operation with average pooling.

The generated stereo volumes in camera frustum space exist the issue of object distortion. To solve the issue, we introduce a multi-scale BEV projection and fusion (MPF) module to

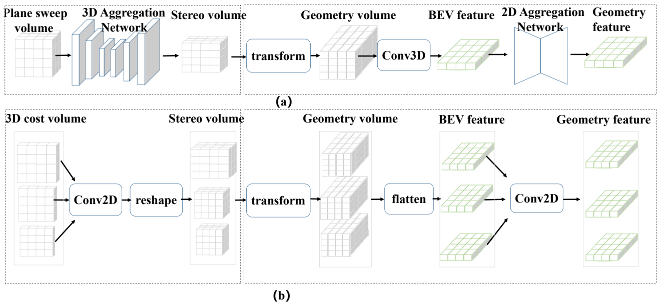


Fig. 3. A detailed comparison between DSGN (a) and our EGFG (b). Our EGFG adopts deep multi-scale information fusion for stereo volume generation (left) and geometry-aware feature generation (right) which avoids complex 3D and 2D aggregation networks adopted in DSGN [11]. Note that, three input layers in our EGFG do not share Conv2D parameters. To highlight the difference to DSGN, we just plot one Conv2D here.

convert stereo volume features in camera frustum space to 3D geometry-aware features in 3D world space.

2) Multi-Scale BEV Projection and Fusion (MPF): As shown in Fig. 2(c), the MPF module first transforms multi-scale stereo volumes ($F^i_{sv}, i = 1, 2, 3$) in camera frustum space to multiple geometry volumes ($F^i_{gv}, i = 1, 2, 3$) in 3D world space. After that, the MPF module converts geometry volumes in 3D space to the features in BEV, and performs a multi-level fusion to generate geometry-aware feature ($F^i_{gf}, i = 1, 2, 3$) for 3D prediction.

To transform stereo volume in camera frustum space to geometry volume in 3D space, we adopt volume transformation operation introduced in DSGN [11]. Specifically, we first generate a regular voxel gird in 3D space and project each voxel in grid into camera frustum space with camera internal parameters. After that, we perform a reversing 3D projection to project the corresponding feature in stereo volume to that in geometry volume. With multi-scale stereo volumes ($F^i_{sv}, i = 1, 2, 3$), we adopt the voxel gird with the same size to generate multiple geometry volumes ($F^i_{gv}, i = 1, 2, 3$) with the same resolution. After that, we convert the geometry volumes to the BEV features ($F^i_{bev} \in \mathbb{R}^{(C*Y)\times X\times Z}, i = 1, 2, 3$) by flattening geometry volumes along channel and y dimensions.

With the BEV features ($F^i_{bev}, i = 1, 2, 3$), we perform a multi-level fusion to generate the enhanced geometry-aware features ($F^i_{gf} \in \mathbb{R}^{C'\times X\times Z}, i = 1, 2, 3$). Specifically, we first use a 2D convolution to generate geometry-aware feature F^1_{gf}. At the same time, we concatenate F^1_{gf} with F^2_{bev} and fed it to a 2D convolution to generate geometry-aware feature F^2_{gf}. Similarly, we generate geometry-ware feature F^3_{gf}. We also map semantic features F^3_l to 3D space and concatenate it with geometry-aware feature F^3_{gf} like [11]. Finally, the geometry-aware feature F^3_{gf} is used to perform 3D prediction.

Difference to DSGN Fig. 3 gives a comparison between DSGN [11] and our EGFG. Both DSGN and our EGFG extract stereo volume in camera frustum space and convert camera frustum space to 3D and BEV spaces for geometry-aware feature generation. Though they adopt a similar pipeline, they have significant differences as follows: (1) The goal is different. DSGN aims to explore an accurate 3D detector without considering computational costs, while our EGFG aims to explore a fast stereo 3D object detector with a best trade-off between speed and accuracy. (2) The key idea of geometry-aware feature generation is different. DSGN employs the heavy 3D and 2D aggregation networks to extract discriminative geometry-aware features, while our EGFG adopts deep multi-scale fusion with 2D convolution to generate discriminative geometry-aware features. The flops of 3D conv in 3D aggregation networks are about $D*K$ times slower than 2D Conv in our EGFG, where D is disparity maximum and K is kernel size. In addition, the number of 3D conv in 3D aggregation networks is lager than the number of 2D conv in EGFG. (4) We observe that it achieves a poor performance without the heavy 3D and 2D aggregation networks used in DSGN [11] (see Table 4).

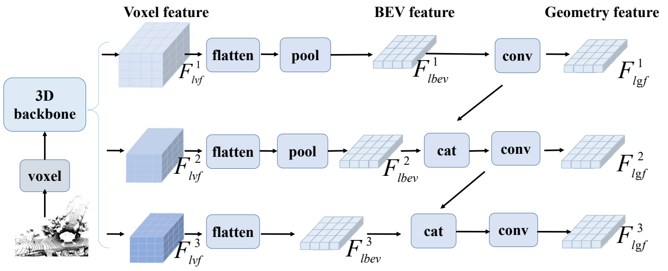


Fig. 4. Detailed architecture of our designed LiDAR detector for distillation. The LiDAR detector first generates multi-scale voxel feature maps by a voxel operation and a sparse 3D convolution backbone like [50]. Then, these multi-scale voxel features are used to generate multiple geometry-aware features.

B. Deep Geometry-Aware Feature Distillation (DGFD)

LiDAR-based 3D detection has a higher accuracy than stereo 3D detection. To bring this gap, Guo *et al.* [15] proposed a novel feature distillation approach (LIGA) for stereo 3d detection. LIGA attaches a single-level feature distillation on the output geometry-aware feature of deep stereo geometry network (DSGN) [11]. We argue that this single-level distillation can not provide a deep supervision. To this end, we propose a deep geometry-aware feature distillation (DGFD) scheme, where a LiDAR-based 3D detector is designed to generate multi-level LiDAR features and then provide deep multi-level supervision for stereo feature learning.

Fig. 4 shows the architecture of our designed LiDAR 3D detector. We first convert raw point cloud representation into a voxel representation and use a spare 3D convolutional backbone [50] to extract multi-scale LiDAR voxel features ($F^i_{lvf}, i = 1, 2, 3$). Then, we flatten LiDAR voxel features along channel and y dimensions and pool the BEV features to the same resolution by using average pooling. We call the resized output features as LiDAR BEV features ($F^i_{lbev}, i = 1, 2, 3$). After that, we perform a multi-level fusion to output LiDAR geometry-aware features ($F^i_{lgf}, i = 1, 2, 3$). Finally, we perform LiDAR 3D prediction on F^3_{lgf}.

We first train this LiDAR 3D detector on point cloud data in KITTI dataset [14]. After that, the LiDAR geometry-aware features ($F^i_{lgf} \in \mathbb{R}^{C'\times X\times Z}, i = 1, 2, 3$) are used to guide stereo geometry-aware feature learning. Specifically, we minimize the feature difference L_{dif} between LiDAR

TABLE I
COMPARISON (AP_{3D}) OF SOME STATE-OF-THE-ART 3D STEREO OBJECT (CAR) DETECTION METHODS ON BOTH KITTI VALIDATION AND TEST SETS. THE INFERENCE TIME, EXCEPT YOLOSTEREO3D, IS TAKEN FROM THE LEADERBOARDS ON OFFICIAL KITTI WEBSITE. YOLOSTEREO3D AND OUR ESGN ARE REPORTED ON NVIDIA RTX3090.

Method	Time	**mAP (test)**			IoU $>$ 0.7 (validation)		
		Moderate	Easy	Hard	**Moderate**	Easy	Hard
TL-Net [35]	-	4.37	7.64	3.74	14.26	18.15	13.72
Stereo RCNN [24]	300ms	30.23	47.58	23.72	36.69	54.11	31.07
IDA3D [30]	300ms	29.32	45.09	23.13	37.45	54.97	32.23
PL: F-PointNet [47]	400ms	26.70	39.70	22.30	39.8	59.4	33.5
PL: AVOD [47]	400ms	34.05	54.53	28.25	45.3	61.9	39
PL++: AVOD [54]	400ms	-	-	-	46.8	63.2	39.8
PL++: P-RCNN [54]	400ms	42.43	61.11	36.99	44.9	62.3	41.6
OC-Stereo [31]	350ms	37.60	55.15	30.25	48.34	64.07	40.39
ZoomNet [49]	300ms	38.64	55.98	30.97	50.47	62.96	43.63
Disp R-CNN [45]	387ms	45.78	68.21	37.73	47.73	64.29	40.11
DSGN [11]	670ms	52.18	73.50	45.14	54.27	72.31	47.71
CG-Stereo [23]	570ms	53.58	74.39	46.50	57.82	76.17	54.63
LIGA [15]	400ms	64.66	81.39	57.22	67.06	84.92	63.80
SNVC [26]	1000ms	61.34	78.54	54.23	63.75	77.29	56.81
RT3DStereo [20]	80ms	23.28	29.90	18.96	-	-	-
Stereo-Centernet [42]	40ms	31.30	49.94	25.62	41.44	55.25	35.13
RTS3D [25]	39ms	37.38	58.51	**31.12**	44.5	63.65	**37.48**
RT3D-GMP [21]	60ms	38.76	45.79	30.00	-	-	-
YOLOStereo3D [27]	50ms	**41.25**	**65.68**	30.42	**46.58**	**72.06**	35.53
ESGN (Ours)	62ms	**46.39(+5.14)**	**65.80(+0.12)**	**38.42(+7.30)**	**52.33(+5.75)**	**72.44(+0.38)**	**43.74(+6.26)**

geometry-aware features and stereo geometry-aware features at multiple levels as follows:

$$L_{dif} = \sum_{i=1,2,3} \frac{1}{N} \left| M_{fg} M_{sp} (g(F^i_{gf}) - F^i_{lgf}) \right|^2, \quad (3)$$

where i represents the scale index, g represents a single 1×1 convolution, M_{fg} is the foreground mask, M_{sp} is LiDAR sparse mask, and N is the number of sparse foreground mask.

C. Prediction Heads

3D Detection Head Similar to [11], [15], we perform 3D detection by a classification head and a regression head. During the training, the total loss of 3D detection can be written as $L_{3d} = L_{cls} + L_{l1} + L_{dir} + L_{iou}$, where L_{cls}, L_{l1}, L_{dir} respectively represent classification loss, box regression loss, and direction classification loss in [15], [50], and L_{iou} represents the rotated IoU loss [61].

Auxiliary Heads Similar to [11], [15], we add two auxiliary heads during training, including a deep estimation head and a 2D detection head. The depth estimation head consists of several convolutions and is attached on stereo volume F^3_{sv} in Fig. 2(b). The ground-truth of depth estimation is transformed from point cloud data. The 2D detection head is attached on the backbone feature F^1_l in Fig. 2(a). During the training, the auxiliary loss can be written as $L_{aux} = L_{depth} + L_{2d}$, where L_{depth} represents depth estimation loss and L_{2d} represents 2D object detection loss.

IV. EXPERIMENTS

A. Dataset and Implementation Details

Dataset We perform the experiments on the classical KITTI dataset [14]. The KITTI dataset consists of 7,481 training paired images and 7,518 test paired images. In addition, the dataset provides the LiDAR point cloud data for each RGB image. Following the existing works [24], [31], [47], we split the original training images into the training set and the validation set. The training set has 3,712 paired images and the validation set has 3,769 paired images. For ablation study, we train our ESGN on the training set with 3,712 images and evaluate it on the validation set. For state-of-the-art comparison, we train our ESGN on the original training images with 7,481 images and submit the results of the test set to the official evaluation server for performance evaluation.

Implementation Details We implement our proposed ESGN on a single NVIDIA RTX3090 GPU. To generate LiDAR features for deep distillation, we first train our designed LiDAR-based detector on the training set using LiDAR point cloud data. We adopt Adam for optimization and set the batch size as 2. There are 80 epochs, where the learning rate is set as 0.003 and decreases at epoch 35 and 45. After that, we train our ESGN on the training set using stereo images. We adopt Adam for optimization and set the batch size as 1. There are 55 epochs, where the learning rate is set as 0.001 and decreases at epoch 50 by a factor of 10. We will plan to implement it with MindSpore in future work.

For KITTI dataset, we set the detection region in range [-30, 30] × [-1, 3] × [2, 59.6] (meters). The voxel size in stereo is set as [0.4m, 0.8m, 0.4m] for regular voxel grid generation (see Sec. III-A1), while the voxel size in LiDAR is set as [0.05m, 0.1m, 0.05m] for LiDAR voxel representation.

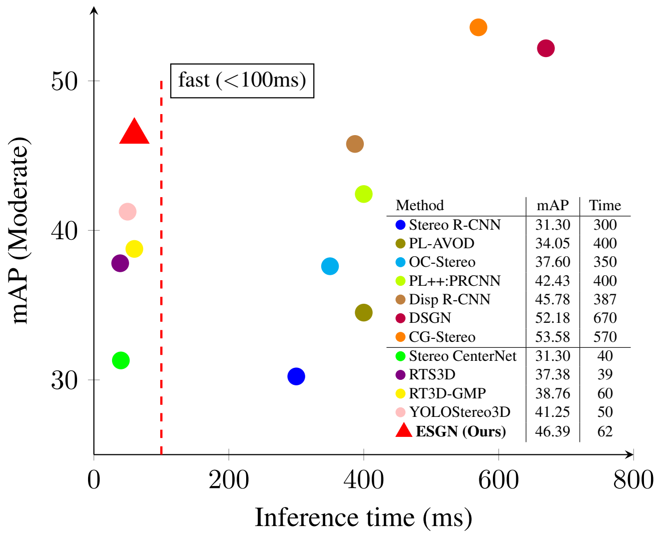
B. Comparison With State-Of-The-Art Methods

We first compare our proposed ESGN with some state-of-the-art methods on both KITTI test and validation sets. According to the degree of occlusion and truncation, the validation and test sets are respectively divided into three subsets: `easy`, `moderate` and `hard`. Tab. I gives the state-of-the-art comparison in terms of speed and AP_{3d}. Compared to the high-accuracy DSGN [11] and LIGA [15], our proposed ESGN is 11.2 and 6.5 times faster in speed. Among these state-of-the-art methods, RT3DStereo [20], Stereo-Centernet [42], RTS3D [25], RT3D-GMP [21], and YOLOStereo3D [27] belong to fast stereo 3D object detection approaches, which have

TABLE II

COMPARISON (AP_{BEV}) OF STATE-OF-THE-ART 3D OBJECT (CAR) DETECTION METHODS ON KITTI VALIDATION SET. THE INFERENCE TIME, EXCEPT YOLOSTEREO3D, IS FROM THE OFFICIAL KITTI LEADERBOARDS. YOLOSTEREO3D AND OUR ESGN ARE REPORTED ON NVIDIA RTX3090.

Method	Time	IoU $>$ 0.7			IoU $>$ 0.5		
		Moderate	Easy	Hard	**Moderate**	Easy	Hard
TL-Net [35]	-	21.88	29.22	18.83	45.99	62.46	41.92
Stereo RCNN [24]	300ms	48.30	68.50	41.47	74.11	87.13	58.93
IDA3D [30]	300ms	50.21	70.68	42.93	76.69	88.05	67.29
PL: F-PointNet [47]	400ms	51.8	72.8	44	77.6	89.8	68.2
PL: AVOD [47]	510ms	39.2	60.7	37	65.1	76.8	56.6
PL++: AVOD [54]	400ms	56.8	74.9	49	77.5	89	68.7
PL++: PIXOR [54]	400ms	61.1	79.7	54.5	75.2	89.9	67.3
PL++: P-RCNN [54]	400ms	56	73.4	52.7	76.6	88.4	69
OC-Stereo [31]	350ms	65.95	77.66	51.20	80.63	90.01	71.06
ZoomNet [49]	300ms	66.19	78.68	57.60	88.40	90.62	71.44
Disp R-CNN [45]	387ms	64.38	77.63	50.68	80.45	90.67	71.03
DSGN [11]	670ms	63.91	83.24	57.83	-	-	-
CG-Stereo [23]	570ms	68.69	87.31	65.80	88.58	97.04	80.34
LIGA [15]	400ms	77.26	89.35	69.05	90.27	97.22	88.36
SNVC [26]	1000ms	72.95	87.07	56.81	-	-	-
Stereo-Centernet [42]	40ms	53.27	71.26	45.53	-	-	-
RTS3D [25]	39ms	**56.46**	76.56	**48.20**	78.70	90.41	**70.03**
YOLOStereo3D [27]	50ms	55.22	**80.69**	43.47	**79.62**	**96.52**	62.50
ESGN (Ours)	62ms	**63.86(+7.4)**	**82.29(+1.6)**	**54.63(+6.43)**	**82.22(+2.6)**	**93.05**	**72.25(+2.22)**



Method	mAP	Time
● Stereo R-CNN	31.30	300
● PL-AVOD	34.05	400
● OC-Stereo	37.60	350
● PL++:PRCNN	42.43	400
● Disp R-CNN	45.78	387
● DSGN	52.18	670
● CG-Stereo	53.58	570
● Stereo CenterNet	31.30	40
● RTS3D	37.38	39
● RT3D-GMP	38.76	60
● YOLOStereo3D	41.25	50
▲ **ESGN (Ours)**	46.39	62

Fig. 5. Accuracy (mAP) and inference time (ms) comparison of some state-of-the-art stereo 3D object detection methods on KITTI test (moderate) set [14]. The inference time of most methods, except YOLOStereo3D [27], are taken from the official KITTI leaderboards. For a fair comparison, the inference time of YOLOStereo3D and our ESGN are reported on a single NVIDIA RTX3090. Our ESGN achieves a best trade-off between accuracy and speed.

TABLE III

IMPACT OF INTEGRATING EGFG (SEC. III-A) AND DGFD (SEC. III-B) MODULES INTO THE BASELINE ON KITTI VALIDATION SET.

Baseline	EGFG	DGFD	**Moderate**	Easy	Hard
✓			15.54	23.89	13.32
✓	✓		49.69	68.05	41.40
✓	✓	✓	52.33	72.44	43.74

the speed of less than $100ms$. For example, YOLOStereo3D [27] achieves an AP_{3d} of 41.25% on moderate test set at the speed of $50ms$. Compared to these fast stereo 3D detection approaches, our ESGN achieves the best accuracy on three subsets of both test and validation sets. For example, our ESGN outperforms YOLOStereo3D by an absolute gain of 5.14% on KITTI moderate test set at a comparable speed.

Tab. II further provides the state-of-the-art comparison in terms of both speed and AP_{bev} on KITTI validation set. We show the results under two evaluation metrics (*i.e.,* IoU $>$ 0.5 and IoU $>$ 0.7). On the moderate set, our ESGN outperforms these fast stereo object detection approaches under these two evaluation metrics. Moreover, we observe that our ESGN is much better under the stricter evaluation metric (IoU $>$ 0.7). For example, our proposed ESGN outperforms YOLOStereo3D by an absolute gain of 8.64% on moderate set with the evaluation metric of IoU $>$ 0.7.

Fig. 5 compares the accuracy and inference time of some stereo 3D object detection methods. Our proposed ESGN has a better trade-off between accuracy and speed than these existing fast stereo methods.

C. Ablation Study

We conduct the ablation study to demonstrate the effectiveness of different modules in our ESGN. All the results in this subsection are evaluated on KITTI validation set under the evaluation metric of IoU $>$ 0.7.

We first show the impact of progressively integrating different modules, including EGFG in Sec. III-A and DGFD in Sec. III-B, into the baseline. Tab. III shows the results on three subsets. Our baseline directly adopts the single stereo volume F_{sv}^1 for stereo 3D detection using 3D head like [27]. On moderate subset, our baseline achieves a very low AP_{3d} of 15.54% on moderate subset, due to very simple and plain design. Then, we integrate our EGFG into the baseline, where EGFG fuses deep multi-level information for both stereo volume and geometry-aware feature generations. Our EGFG achieves an AP_{3d} of 49.69%, which demonstrates the effectiveness of our EGFG module. Finally, we integrate our DGFD into them, which achieves an AP_{3d} of 52.33%. It demonstrates that deep distillation is useful for 3D detection.

We further show the impact of different modules in our EGFG in Tab. IV. Our EGFG contains a SCR module and a MPF module. When using only single scale SCR (baseline

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3202810

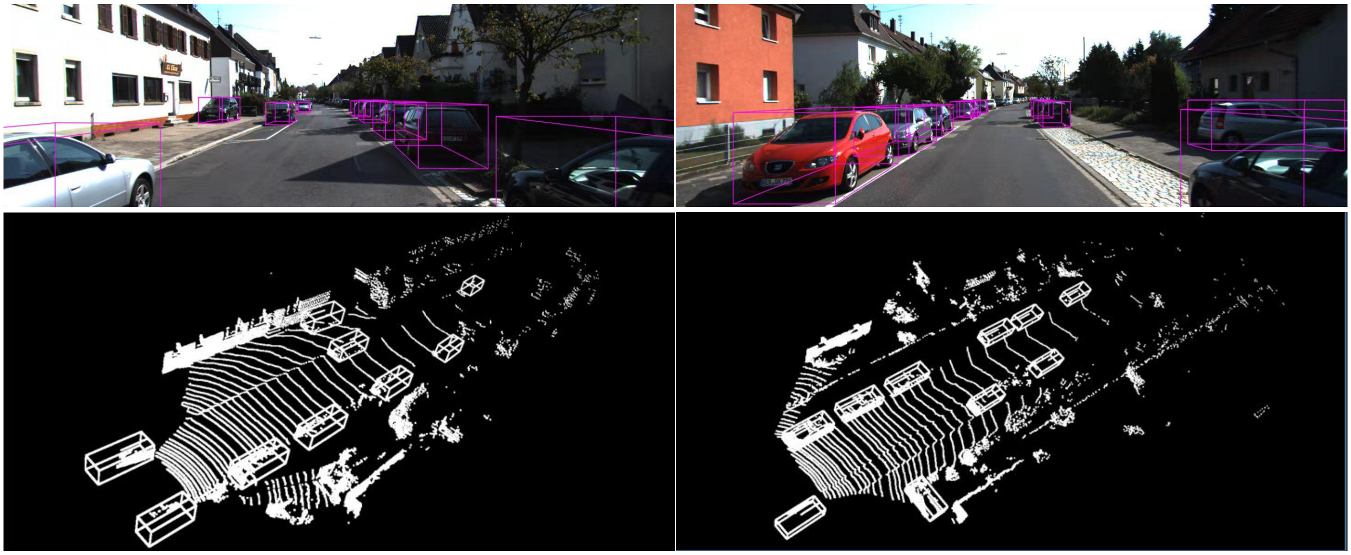


Fig. 6. Qualitative stereo 3D results of our ESGN. The top row shows 3D detection results in image viewpoint, and the bottom row shows 3D detection results in 3D space. Our ESGN detects both large and small objects accurately.

TABLE IV
IMPACT OF TWO MODULES, INCLUDING SCR (SEC. III-A1) AND MPF (SEC. III-A2), IN OUR EGFG MODULE (SEC. III-A) ON KITTI VALIDATION SET. 'SINGLE' REPRESENTS USING ONLY ONE OF THREE SCALES WITHOUT MULTI-SCALE FUSION, WHILE 'MULTIPLE' REPRESENTS USING THREE SCALES WITH MULTI-SCALE FUSION.

SCR		MPF		**Moderate**	Easy	Hard
Single	Multiple	Single	Multiple			
✓				15.54	23.89	13.32
✓		✓		37.87	57.12	31.52
	✓	✓		46.24	64.98	38.79
	✓		✓	49.69	68.05	41.40

TABLE V
IMPACT OF THE NUMBER OF LAYERS FOR MULTI-SCALE FUSION IN OUR EGFG ON KITTI VALIDATION SET.

Method	**Moderate**	Easy	Hard
Single	37.87	57.12	31.52
Two	44.41	63.93	36.71
Three	49.69	68.05	41.40

in Tab. III), it achieves an AP_{3d} of 15.54% on moderate subset. When using single-scale SCR and single-scale MPF, it achieves an AP_{3d} of 37.87%. Note that, this single-scale SCR and single-scale MPF setting is similar to the light-weighted DSGN in which the heavy 3D and 2D aggregation networks are removed. The light-weighted DSGN is inferior to our ESGN. When using multi-scale SCR and single-scale MPF, it achieves an AP_{3d} of 46.24%, which provides 8.37% improvement. Namely, single-scale SCR can not extract multi-scale geometry information for stereo volume generation. Based on multi-scale SCR, we perform multi-scale MPF and further improve the performance by 3.45%. Namely, multi-scale BEV projection and fusion further enhances geometry-aware features. In addition, we show the impact of the number of layers for multi-scale fusion in our EGFG in Tab. V. Compared to single layer or two-layer fusion, three-layer fusion has the best performance.

TABLE VI
IMPACT OF SINGLE-LEVEL DISTILLATION AND DEEP DISTILLATION IN OUR DGFD MODULE (SEC. III-B) ON KITTI VALIDATION SET.

Method	**Moderate**	Easy	Hard
No distillation	49.69	68.05	41.40
Single-level distillation	51.40	72.25	43.31
Deep distillation	52.33	72.44	43.74

We also compare single-level distillation and deep distillation in Tab. VI. Single-level distillation only guides feature learning at single geometry-aware feature F_{gf}^3 like [15], while deep distillation guides feature learning at multiple geometry-aware features $F_{gf}^i, i = 1, 2, 3$. Compared to no distillation design, these two distillation strategies respectively provide 1.71% and 2.64% improvements on moderate subset. Compared to single-level distillation, our proposed deep distillation has 0.93% improvement. It demonstrates that deep multi-level distillation can better guide geometry-aware feature learning in multiple levels. In addition, our deep distillation does not add extra computation costs during inference.

Finally, we provide some qualitative 3D detection results of our ESGN in Fig. 6. The results in image viewpoint (top) and corresponding 3D space (bottom) are both provided. Our ESGN can detect both large and small objects accurately.

V. CONCLUSION

In this paper, we have proposed an efficient stereo geometry network (ESGN) for fast 3D object detection. The key module is a novel efficient geometry-aware feature generation (EGFG) module that first generates multi-scale stereo volumes by a SCR module and second generates geometry-aware features by a MPF module. With deep multi-scale fusion, our EGFG module generates discriminative geometry-aware features without heavy aggregation operation. We also introduce a deep geometry-aware feature distillation scheme to guide feature learning with a LiDAR-based 3D detector. We perform experiments on KITTI dataset. Our ESGN achieves a best trade-off between speed and accuracy. Compared to the high-accuracy DSGN, our ESGN is 11.2 times faster in speed. Compared to the fast YOLOStereo3D, our ESGN achieves an AP_{3d} improvement of 5.14% at a comparable fast speed. We hope that our ESGN can provide more possible ways for fast stereo 3D object detection.

References

[1] Libo Sun, Haokui Zhang, and Wei Yin. Pseudo-lidar based road detection. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5386-5398, 2022. 2

[2] Zhimin Lu, Jue Wang, Zhiwei Li, Song Chen, and Feng Wu. A resource-efficient pipelined architecture for real-time semi-global stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 660-673, 2022. 2

[3] Lichen Zhao, Jinyang Guo, Dong Xu, and Lu Sheng. Transformer3d-det: Improving 3d object detection by vote refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4735-4746, 2021. 3

[4] Chenglong Xu, Chengdong Wu, Daokui Qu, Fang Xu, Haibo Sun, and Jilai Song. Accurate and efficient stereo matching by log-angle and pyramid-tree. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4007-4019, 2021. 2

[5] Lincheng Li, Shunli Zhang, Xin Yu, and Li Zhang. Pmsc: Patchmatch-based superpixel cut for accurate stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 679-692, 2018. 2

[6] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[7] Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. Hierarchical shot detector. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 9705-9714. 2

[8] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 568-579, 2018. 3

[9] Runmin Cong, Haowei Yang, Qiuping Jiang, Wei Gao, Haisheng Li, Cong Wang, Yao Zhao, and Sam Kwong. Bcs-net: Boundary, context and semantic for automatic covid-19 lung infection segmentation from CT Images. *IEEE Transactions on Instrumentation and Measurement*, 2022. 3

[10] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 9774-9783. 3

[11] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12533-12542. 1, 2, 3, 5, 6, 7

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 3

[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 6569-6578. 2

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354-3361. 2, 5, 6, 7

[15] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. *Proc. IEEE International Conference on Computer Vision*, 2021, pp. 3313-3343. 1, 2, 3, 5, 6, 7, 8

[16] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11870-11879. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778. 3, 4

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 1921-1930. 3

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 3

[20] Hendrik Königshof, Niels Ole Salscheider, and Christoph Stiller. Realtime 3d object detection for automated driving using stereo vision and semantic information. *Proc. International IEEE Conference on Intelligent Transportation Systems*, 2019, pp. 1405-1410. 1, 3, 6

[21] Hendrik Königshof and Christoph Stiller. Learning-based shape estimation with grid map patches for realtime 3d object detection for automated driving. *Proc. International IEEE Conference on Intelligent Transportation Systems*, 2020, pp. 1-6. 1, 3, 6

[22] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12697-12705. 3

[23] Chengyao Li, Jason Ku, and Steven L. Waslander. Confidence guided stereo 3d object detection with split depth estimation. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 5776-5783. 6, 7

[24] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7636-7644. 1, 2, 3, 6, 7

[25] Peixuan Li, Shun Su, and Huaici Zhao. Rts3d: Real-time stereo 3d detection from 4d feature-consistency embedding space for autonomous driving. *Proc. AAAI Conference on Artificial Intelligence*, 2021, pp. 1930-1939. 6, 7

[26] Shichao Li, Zechun Liu, Zhiqiang Shen, and Kwang-Ting Cheng Stereo neural vernier caliper. *Proc. AAAI Conference on Artificial Intelligence*, 2022, pp. 1376-1385. 6, 7

[27] Yuxuan Liu, Lujia Wang, and Ming Liu. Yolostereo3d: A step back to 2d for efficient stereo 3d detection. *Proc. International Conference on Robotics and Automation*, 2021, pp. 13018-13024. 1, 2, 3, 6, 7

[28] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695-5703. 4

[29] Timo Ojala, Matti PietikaEinen, and Topi MaEenpaEa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2004. 3

[30] Wanli Peng, Hao Pan, He Liu, and Yi Sun. Ida-3d: Instance-depth-aware 3d object detection from stereo vision for autonomous driving. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13015-13024. 1, 2, 6, 7

[31] Alex D. Pon, Jason Ku, Chengyao Li, and Steven L. Waslander. Object-centric stereo matching for 3d object detection. *Proc. IEEE International Conference on Robotics and Automation*, 2020, pp. 8383-8389. 2, 6, 7

[32] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 9277-9286. 3

[33] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652-660. 1, 3

[34] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Proc. International Conference on Neural Information Processing Systems*, 2017, pp. 5099-5108. 1, 3

[35] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: From monocular to stereo 3d object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7607-7615. 6, 7

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. International Conference on Neural Information Processing Systems*, 2015, pp. 91-99. 2

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *Proc. The International Conference on Learning Representations*, 2015. 3

[38] 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. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10526-10535. 3

[39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. *Proc. IEEE*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 770-779. 1, 3

[40] 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. no. 8, pp. 2647-2664, 2021. 3

[41] Weijing Shi and Ragunathan (Raj)Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711-1719. 3

[42] Yuguang Shi, Yu Guo, Zhenqiang Mi, and Xinjie Li. Stereo centernet based 3d object detection for autonomous driving. *arXiv:2103.11071*, 2021. 3, 6, 7

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 3

[44] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[45] Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qinhong Jiang, Xiaowei Zhou, and Hujun Bao. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10548-10557. 2, 6, 7

[46] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet: A fast and accurate network for disparity estimation. *arXiv:2003.10758*, 2020. 4

[47] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8437-8445. 1, 2, 6, 7

[48] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10447-10456. 3

[49] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. *Proc. AAAI Conference on Artificial Intelligence*, 2020, pp. 12557-12564. 2, 6, 7

[50] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, vol. 18, no. 10, pp. 1-17, 2018. 5, 6

[51] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11040-11048. 3

[52] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse to-dense 3d object detector for point cloud. *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 1951-1960. 3

[53] Maosheng Ye, Shuangjie Xu, and Tongyi Cao. Hvnet: Hybrid voxel network for lidar based 3d object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1631-1640. 3

[54] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *Proc. International Conference on Learning Representations*, 2020. 2, 6, 7

[55] Jun Yu, Jinghan Yao, Jian Zhang, Zhou Yu, and Dacheng Tao. Sprnet: Single-pixel reconstruction for one-stage instance segmentation. *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1731-1742, 2021. 3

[56] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[57] Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection. *Proc. International Conference on Neural Information Processing Systems*, 2020, pp. 6959-6970. 3

[58] Jian Zhang, Yunyin Cao, Qun Wu. Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recognition*, vol. 116, 2021. 3

[59] Jian Zhang, Jun Yu, and Dacheng Tao. Local deep-feature alignment for unsupervised dimension reduction. *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2420-2432, 2018. 3

[60] Jian Zhang, Jianing Yang, Jun Yu, and Jianping Fan. Semisupervised image classification by mutual learning of multiple self-supervised models. *International Journal of Intelligent Systems*, vol. 27, no. 5, pp. 2420-2432, 2022. 3

[61] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. *Proc. International Conference on 3D Vision*, 2019, pp. 85-94. 6

[62] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490-4499. 3

Aqi Gao received the B.S. degree in Electronic information engineering from Hefei University of Technology, China, in 2020. And he is currently pursuing the master degree in information and communication engineering from Tianjin University, Tianjin, China, under the supervision of Prof. Y. Pang. His research interests is 3D object detection of stereo, monocular and LiDAR.

Yanwei Pang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China in 2004. He is currently a Professor with Tianjin University, China, where he is also the Founding Director of the Tianjin Key Laboratory of Brain Inspired Intelligence Technology (BIIT). His research interests include object detection and image recognition, in which he has published 150 scientific papers, including 40 IEEE Transactions and 30 top conferences (e.g., CVPR, ICCV, and ECCV) papers.

Jing Nie (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from Tianjin University, China, in 2017, where she is currently pursuing the Ph.D. degree under the supervision of Prof. Y. Pang. Her research interests include object detection and image dehazing.

Zhuang Shao received the B.Eng. degree in electronic and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2015, and the M.Sc. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2018. He is currently pursuing the Ph.D. degree with the Warwick Manufacturing Group, University of Warwick, Coventry, U.K. His research interests include image captioning, video captioning, and machine learning

Jiale Cao received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2018. He is currently an Associate Professor with Tianjin University. His research interests include object detection and image analysis, in which he has published 20+ IEEE Transactions and CVPR/ICCV/ECCV articles. He serves as a regular Program Committee Member for leading computer vision and artificial intelligence conferences, such as CVPR, ICCV, and ECCV.

Yishun Guo received the B.S. degree in Internet of things engineering from Tianjin University, Tianjin, China, in 2022. And he is currently pursuing the master degree in information and communication engineering in Tianjin University, Tianjin, China, under the supervision of Prof. Y. Pang. His research interests is object detection.

Xuelong Li (Fellow, IEEE) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.