Monocular 3D Object Detection with Depth from Motion

Tai Wang[1,2] Jiangmiao Pang[2] Dahua Lin[1,2]

[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory
{wt019,dhlin}@ie.cuhk.edu.hk, pangjiangmiao@gmail.com

Abstract. Perceiving 3D objects from monocular inputs is crucial for robotic systems, given its economy compared to multi-sensor settings. It is notably difficult as a single image can not provide any clues for predicting absolute depth values. Motivated by binocular methods for 3D object detection, we take advantage of the strong geometry structure provided by camera ego-motion for accurate object depth estimation and detection. We first make a theoretical analysis on this general two-view case and notice two challenges: 1) Cumulative errors from multiple estimations that make the direct prediction intractable; 2) Inherent dilemmas caused by static cameras and matching ambiguity. Accordingly, we establish the stereo correspondence with a geometry-aware cost volume as the alternative for depth estimation and further compensate it with monocular understanding to address the second problem. Our framework, named Depth from Motion (DfM), then uses the established geometry to lift 2D image features to the 3D space and detects 3D objects thereon. We also present a pose-free DfM to make it usable when the camera pose is unavailable. Our framework outperforms state-of-the-art methods by a large margin on the KITTI benchmark. Detailed quantitative and qualitative analyses also validate our theoretical conclusions. The code is released at https://github.com/Tai-Wang/Depth-from-Motion.

Keywords: Monocular 3D Object Detection, Depth from Motion

1 Introduction

3D object detection is a fundamental task for practical applications such as autonomous driving. In the past few years, LiDAR-based [16,42,45,28] and binocular-based [7,9,17,5,12] approaches have made great progress and achieved promising performance. In contrast, monocular methods [37,23,30,35] still yield unsatisfactory results as their depth estimation is naturally ill-posed. Although several works [23,30,35,44,29] made some attempts to tackle this problem, the current solutions still focus on digging out more geometry structures from *a single image*. It is still hard for them to estimate accurate *absolute* depth values.

This paper aims to use stereo geometry from a pair of images nearby in temporal to facilitate the object depth estimation. The basic principle is similar to depth estimation in binocular systems. Two cameras in binocular systems are strictly constrained on the same plane and have a fixed distance, which

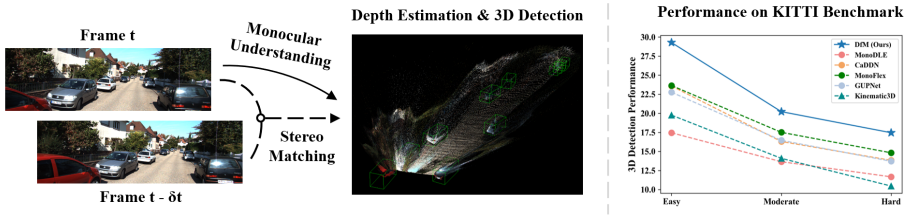

Fig. 1. In this paper, we present a framework for monocular 3D detection from videos. In contrast to previous work only relying on monocular understanding from a single image, our method integrates the stereo geometric clues from temporally adjacent images. It significantly improves depth estimation accuracy, the most critical part for camera-only 3D perception, and thus enhances the 3D detection performance.

is known as the system's *baseline*. State-of-the-art stereo 3D object detection methods take this baseline as a critical clue and transform depth estimation to an easier disparity estimation problem. Similarly, two nearby images in temporal also have stereo correspondence, but their baseline is dynamic and relies on the ego-motion of the camera. This idea is intuitively promising, but few previous works explored it. The only recent work for 3D detection from monocular videos, Kinematic3D [3], uses a 3D Kalman Filter and an integrated ego-motion module to build the connection between frames. It focuses on the robustness and stability of detection results but still estimates depth from a single image. Our work, instead, is the first to study how to improve object depth estimation and 3D detection from the strong stereo geometry formed by ego-motion.

We first conduct a theoretical analysis on this problem to better understand the geometry relationship. It reveals that direct derivation of depth in this setting involves many estimations and thus has fundamental difficulty caused by cumulative errors. The stereo estimation also has several intrinsic dilemmas, such as no baseline formed by static cameras. We thus build our framework with a depth-from-motion module addressing these problems to construct 3D features and detect 3D objects thereon. Specifically, we first involve the complex geometry relationship in a differential cost volume as the alternative for stereo estimation. To guarantee its physical rationality for any arbitrarily augmented inputs, we devise a pipeline to ensure the pose transformation takes place in the original space, namely *canonical space*. Furthermore, we compensate it with another monocular pathway and fuse them with learnable weights. The distribution of these learned weights well demonstrates the theoretical discussion on the intrinsic weaknesses of stereo estimation.

Considering camera poses are not always available, we also introduce a pose-free method to make the framework more flexible. We first decouple the ego-pose estimation as translation and rotation. Instead of using the straightforward Euler angles, we formulate the rotation with quaternion, a more friendly representation for optimization, to avoid periodic targets. In addition, we adopt a self-supervised

loss to regularize the learning of pose to make the training get rid of pose annotations and expensive loss weights tuning.

We evaluate our framework on the KITTI [8] benchmark. It achieves 1st place out of monocular methods, surpassing previous methods by a large margin, 2.6%∼5.6% and 4.2%∼7.5% AP higher on the 3D and bird-eye-view vehicle detection benchmark respectively. These impressive experimental results demonstrate the potential of this stream of methods in this context, which is a more interpretable and practical perception approach like that human beings rely on.

2 Related Work

Video-Based Depth Estimation Depth estimation from monocular videos is an important problem for mobile devices and VR/AR applications. Learning-based methods can be divided into MVS-based (Multi-View-Stereo) methods [19,34] and monocular-stereo hybrid methods [43,22,15]. The former group can not handle dynamic scenes due to the static assumption of MVS, and the latter addresses this problem by integrating a pretrained single-view depth estimator. In addition, there is another line of work [10,11] using videos as supervision to achieve self-supervised depth estimation. Although these works have made progress in this problem, there is still a notable gap between this field and camera-only 3D detection. Due to the disparity of scenarios and ultimate targets, previous work hardly attempts to tackle the object depth estimation problem in our context.

Video-Based Object Detection Video-based object detection [48,47,39,1,20] has been studied for several years in the 2D case. These works target a better trade-off between accuracy and efficiency by aggregating features from multiple frames. Unlike the 3D case, the main problems of 2D detection from videos are the occlusion and blur of objects. The transformation between frames is generally flow-based, without considering geometric consistency in the real world. In comparison, the only previous work [3] for monocular 3D video object detection improves the robustness of detection results with 3D Kinematic designs. This paper is different from both. We instead focus on the specific problem in the 3D case: estimating object depth more accurately from the depth-from-motion setting and further boosting the 3D detection performance.

Camera-Only 3D Object Detection Compared to LiDAR-based approaches [16,42,45,28,46,36], camera-only methods take RGB images as the only input and need to reason the depth information without accurate measurement provided by depth sensors. Among them, monocular 3D detection is more challenging than binocular because of its ill-posed property.

Earlier learning-based monocular methods [4,40,24] used sub-networks to solve this problem. Afterward, due to the system complexity and dependence on external data and pretrained models, recent work turns to end-to-end designs [2,13,32,23,37] like 2D detection. As several works [35,30,23] point out the crucial role of depth estimation in this setting, a stream of work [18,35,44,29] attempted to address the problem with more geometric designs. Meanwhile, another line incorporates depth information to study the feature transformation

approaches. Pioneer work [38,27] in this line transforms the input image to 3D representations with depth estimation and performs 3D object detection thereon. Recent CaDDN [26] merges these two stages into an end-to-end framework and achieves promising results. Our work follows this high-level pipeline while focusing on improving the depth estimation from video input.

As for binocular methods, apart from the previously mentioned Pseudo-LiDAR fashion, they can be grouped into two tracks: front-view 2D-based [17,33,41,25] and bird-eye-view volume-based [5,12]. The volume-based methods are consistent with the feature transformation ideas of CaDDN. Our framework is also motivated by this stream. In contrast, we focus on studying a more difficult stereo setting: general multi-view cases formed by ego-motion.

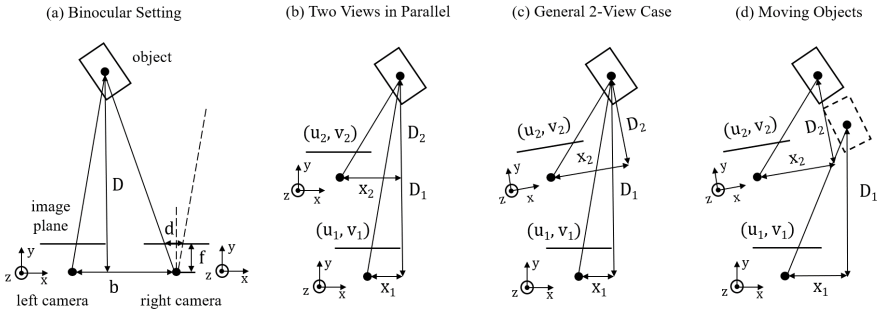

Fig. 2. Multi-view geometry for object depth estimation in the (a) binocular, (b) parallel two-view, (c) general two-view system and (d) that for moving objects.

3 Theoretical Analysis

In this section, we will first make a theoretical analysis for general stereo depth estimation. Among different multi-view settings, the binocular case is the simplest one and thus studied the most in the driving scenario [17,33,5]. We start with this setting and further discuss the connection and difference when extending it to general cases. Finally, we analyze the main challenges in the depth-from-motion setting and introduce our framework design thereon.

3.1 Object Depth from Binocular Systems

Binocular systems strictly constrain two cameras on the same plane. As shown in Fig. 2-(a), the focal length of cameras and the distance between the pair of cameras (namely *baseline* of the system) are supposed to be constant. Following the similar triangle rule under the pinhole camera model, they follow

$$\frac{d}{f} = \frac{b}{D} \Rightarrow D = f\frac{b}{d}, \tag{1}$$

where d is the horizontal disparity on the pair of images, f is the focal length of cameras, D is the object depth, b is the baseline. Following Eq. 1, object depth estimation can be transformed to a much easier disparity estimation problem.

3.2 Object Depth from General Two-View Systems

Binocular systems rely on two-view stereo geometry to estimate object depth. Intuitively, two nearby images in a video also have similar stereo correspondence. Can we use two-view geometry in this general case to predict object depth?

We step by step extend the geometry relationship in binocular systems to general two-view cases. The analysis supposes the camera is in different positions at time t_1 and t_2 respectively, and we know the camera parameters at each position. We assume all objects do not move at the beginning of this analysis and discuss the object motion at the end.

As shown in Fig. 2-(b), suppose the camera's movement only involves translation. We can obtain Δx and ΔD from the transformation of camera poses. The two-view geometry in this parallel case satisfies

$$\frac{u_1 - c_u}{f} = \frac{x_1}{D_1}, \quad \frac{u_2 - c_u}{f} = \frac{x_2}{D_2}, \quad \Delta x = x_1 - x_2, \quad \Delta D = D_1 - D_2, \tag{2}$$

where (u_1, v_1) and (u_2, v_2) are a pair of corresponding points on the images, D_1 and D_2 are their depths, x_1 and x_2 are their locations in 3D space along the x-axis. From these relationships, we can derive D_1:

$$D_1 = \frac{f(\Delta x - \frac{u_2 - c_u}{f}\Delta D)}{u_1 - u_2} \xlongequal{\Delta D = 0} \frac{f\Delta x}{u_1 - u_2}. \tag{3}$$

The geometry relationship in binocular systems is its special case when $\Delta D = 0$.

As Eq. 3 shows, in contrast to binocular system, the "baseline" in this case is no longer fixed but dynamic that relies on camera ego-motion $\Delta x, \Delta D$ and object absolute locations u_2. Accordingly, object depth estimation also relies on them apart from the disparity $u_1 - u_2$.

To better understand this case, we quantitatively compare it with the binocular system on KITTI as an example. It is well-known that a suitable baseline should not be too large or small. A too-large baseline yields small shared regions of two images, while a too-small baseline results in small disparities and large estimation errors. So we take the binocular baseline (0.54 meters on KITTI) as our example target to form with $\Delta x - \frac{u_2 - c_u}{f}\Delta D$ in this case. Because the horizontal translation Δx is typically much smaller than 0.54 meters, we need a large translation along the depth direction (ΔD) and a large horizontal distance from the 2D camera center ($u_2 - c_u$) to get a baseline large enough for stereo matching. For example, to form the 0.54-meter baseline, when ΔD is 5.4 meters, f is 700 pixels, then we need $u_2 - c_u = 70$. Accordingly, when ΔD is only 2.7 meters, then we need $u_2 - c_u = 140$ [1]. It means we can get more accurate estimations for objects far from central lines and may encounter problems otherwise.

[1] For reference, the half-width of an image on KITTI is about 600 pixels.

On this basis, involving ego-rotation (Fig. 2-(c)) will introduce rotation coefficients entangled with object absolute positions to the disparity computation, and involving object motion (Fig. 2-(d)) will introduce *relative* translation and rotation factors. More introduction of absolute positions and motion estimation errors makes direct depth estimation more difficult. See more derivation details in the supplementary materials.

3.3 Achilles Heel of Depth from Motion

Based on the previous analysis, we can observe that direct derivation of depth in a general two-view system involves many estimations like object absolute locations and motions, thus having fundamental difficulties caused by cumulative errors. In addition, the stereo-based solution has several cases that are intrinsically hard to handle, such as no baseline formed by static cameras and the common ambiguity problem of matching on less-textured regions.

Therefore, motivated by binocular approaches [5], we involve the complex geometric relationship in a differential plane-sweep cost volume as the alternative to establish the stereo correspondence: Considering we can not directly estimate depth from disparity, we instead provide candidate depths for each pixel, reproject these 2.5D points to another frame and learn which one is most likely according to the pixel-wise feature similarity. Furthermore, to address the second challenge, we introduce another path for monocular understanding to compensate the stereo estimation. Next, we will elaborate on these designs with our framework in detail.

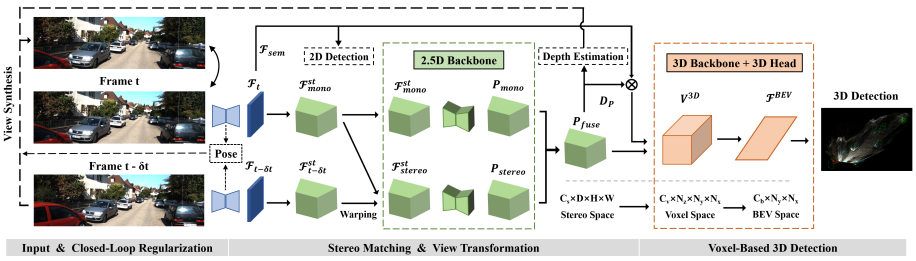

Fig. 3. An overview of our framework.

4 Methodology

A general pipeline for camera-only 3D detection methods typically consists of three stages: extracting features from input images, lifting the features to 3D space, and finally detecting 3D objects thereon. We build our framework following this approach (Fig. 3). Next, we will introduce our overall framework and present two key designs: geometry-aware cost volume construction and

monocular compensation for stereo estimation. Finally, we propose a solution for pose-free cases, making the framework more flexible.

4.1 Framework Overview

2D Feature Extraction Motivated by binocular approaches [5,12], given the input image-pair $(I_t, I_{t-\delta t})$, we first use a shared 2D backbone to extract their features $(\mathcal{F}_t, \mathcal{F}_{t-\delta t})$. Afterward, we devise two different necks to generate F_t as geometric feature for stereo matching and F_{sem} as semantic feature following [12]. To guarantee the semantic features can get correct supervision signals, they are also used to perform the auxiliary 2D detection.

Stereo Matching and View Transformation After getting the features of two frames, we construct the stereo cost volume $\mathcal{F}^{st}_{stereo}$ with the pose transformation between them. In addition, we lift $\mathcal{F}_t$ with pre-defined discrete depth levels to get $\mathcal{F}^{st}_{mono}$ in stereo space for subsequent monocular understanding. A dual-path 3D aggregation network filters these two volumes to predict the depth distribution volume D_P. $D_P(u, v, :)$ represents the depth distribution of pixel (u, v) over the depth levels. The depth prediction is supervised with projected LiDAR points. Details of cost volume construction and the dual-path feature aggregation will be presented in Sec. 4.2 and 4.3. Subsequently, we lift the semantic feature $\mathcal{F}_{sem}$ with D_P, combine it with geometric stereo feature P_{stereo} as the final stereo feature, and sample voxel features thereon. As shown in Fig. 3, this process transforms the feature in stereo space to voxel space, which has a regular structure and is thus more convenient for us to perform object detection.

Voxel-Based 3D Detection Next, we merge the channel dimension and height dimension to transform the 3D feature V^{3D} to bird-eye-view (BEV) space, and apply a 2D hourglass network to aggregate the BEV features. Finally, a lightweight head is appended to predict 3D bounding boxes and their categories. The training loss is composed of two parts as [12]: depth regression loss and 2D/3D detection loss. See more details in the supplemental materials.

4.2 Geometry-Aware Stereo Cost Volume Construction

The key component in the previously mentioned stereo matching is the construction of cost volume. In contrast to the binocular case, the pose transformation between two frames is a rigid transformation composed of translation and rotation. This difference affects the method to construct cost volume and makes it hard to perform data augmentation on input images. Next, we will first formulate the procedure of volume construction and then present how we make it compatible with arbitrarily augmented input.

Formally, for each position $\mathbf{x} = (u, v, w)$ in the stereo volume, we can derive the reprojection matrix $\mathcal{W}$ to warp $\mathcal{F}_{t-\delta t}$ to the space of frame t and concatenate the corresponding feature together:

$$\mathcal{F}^{st}_{stereo}(u_t, v_t, w_t) = concat\Big[\mathcal{F}_t(u_t, v_t), \mathcal{F}_{t-\delta t}(u_{t-\delta t}, v_{t-\delta t})\Big], \tag{4}$$

$$(u_{t-\delta t}, v_{t-\delta t}, d(w_{t-\delta t}))^T = \mathcal{W}(u_t, v_t, d(w_t))^T, \quad \mathcal{W} = KTK^{-1}. \tag{5}$$

Here (u_t, v_t, w_t) and $(u_{t-\delta t}, v_{t-\delta t}, w_{t-\delta t})$ represent the queried pixel coordinates in the stereo space of two frames. $d(w) = w \cdot \Delta d + d_{min}$ is the function to calculate the corresponding depth, where Δd is the divided depth interval and d_{min} is the minimum depth of detection range. $\mathcal{W}$ is the reprojection matrix, which is derived by multiplying intrinsic matrix K, ego-motion (rigid transformation) T and K^{-1}, assuming the intrinsic matrix does not change across two frames. We find that any data augmentation, such as image rescale or flip, can affect the physical rationality of reprojection matrix $\mathcal{W}$. Constructing a *geometry-aware* cost volume from augmented images here is not as trivial as in previous camera-only detection methods. Therefore, we devise an approach to addressing this problem.

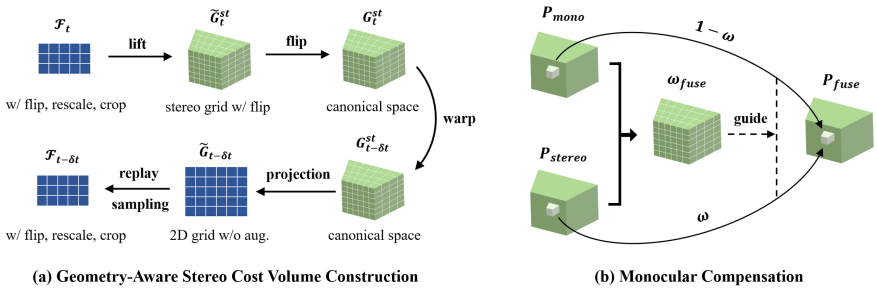

Fig. 4. Key components in our depth-from-motion module.

As shown in Fig. 4-(a), we need to find the corresponding features between a pair of augmented image features $(\mathcal{F}_t, \mathcal{F}_{t-\delta t})$. Our key idea is to guarantee the warping transformation is conducted in the 3D real world, namely *canonical space*. For example, if we perform flipping, rescaling, and cropping on the input two images, we first need to append pre-defined depth levels to each 2D grid coordinate of $\mathcal{F}_t$ and lift each 2.5D coordinate to 3D. During transformation, the effect of *intrinsic* augmentations like rescaling and cropping should be removed through the manipulated[2] intrinsic matrix K. Afterward, we flip the stereo grid $\tilde{G}_t^{st}$ to get G_t^{st} in the canonical space. With the recovered grid, we can further perform the pose transformation to get $G_{t-\delta t}^{st}$, project it to the 2D plane and obtain several $G_{t-\delta t}$ grid maps. Finally, we replay the image augmentations and sample the corresponding features.

In this way, we can exploit any data augmentation to the input images without influencing the intrinsic rationality of ego-motion transformation. Compared to the tricky image swapping for flip augmentation in the binocular case and other alternatives, our method is also generalizable for other multi-view cases.

[2] Rescaling and cropping correspond to the manipulation of focal length and camera centers proportionally.

4.3 Monocular Compensation

The underlying philosophies of stereo and monocular depth estimation are different: stereo estimation relies on matching while monocular estimation relies on the semantic and geometric understanding of a single image and data-driven priors. As analyzed in Sec. 3.3, there are multiple cases that stereo estimation approaches can not handle. Therefore, we incorporate the monocular contextual prior to compensate stereo depth estimation.

Specifically, as shown in Fig. 4-(b), we use two 3D hourglass networks to aggregate monocular and stereo features separately. The network for monocular path shares the same architecture with the other, except for the input channel is half given the $\mathcal{F}^{st}_{mono}$ is half of $\mathcal{F}^{st}_{stereo}$. Then we have two feature volumes P_{mono} and P_{stereo} in the stereo space with the same shape. To aggregate these two features, we devise a simple yet effective and interpretable scheme. First, P_{mono} and P_{stereo} are concatenated and fed into a simple 2D convolutional network composed of 1×1 kernel, and aggregated along the depth channel, *e.g.*, compressed from $2D$ channels to D. Then the sigmoid response of this feature serves as the weight ω_{fuse} for guiding the fusion of P_{mono} and P_{stereo}. Formally, denoting the convolutional network as ϕ, this procedure is represented as follows:

$$\omega_{fuse} = \sigma(\phi(P_{mono}, P_{stereo})), \quad P_{fuse} = \omega_{fuse} \circ P_{stereo} + (1 - \omega_{fuse}) \circ P_{mono} \quad (6)$$

Here σ denotes the sigmoid function, and $\circ$ refers to element-wise multiplication. The derived stereo feature P_{fuse} is directly used to predict the depth distribution after a softmax and also fed into the subsequent networks for 3D detection.

This design is clean yet effective, as to be shown in the ablation studies of Sec. 5.4. Furthermore, it is interpretable both intuitively and empirically. The weight distribution of each position on the image is derived from monocular and stereo depth distributions of the same position. It is location-aware for different regions on the image, agnostic to specific reasons of inaccurate stereo estimation, and self-adaptive to different input cases. We can also validate this expected behavior by visualizing the weight ω_{fuse} and observe where stereo or monocular estimation is more reliable. See more visualization analysis in Sec. 5.3.

4.4 Pose-Free Depth from Motion

Now we have an integrated framework for estimating depth and detecting 3D objects from consecutive-frame images. In the framework, ego-pose serves as a critical clue like the baseline in the binocular case. We essentially estimate the metric-aware depth given the metric-aware pose transformation. Although it can be easily obtained in practical applications, here we still propose a solution for the pose-free case. It is useful for mobile devices in the wild and necessary for evaluating our final models on the KITTI [8] test set.

The target formulation is critical for camera pose estimation. It is well known that any rigid pose transformation can be decomposed as translation

and rotation. Both have three Degrees of Freedom (DoF). Previous work [3,10] typically regresses the 3D translation and three Euler angles. The regression of translation $\mathbf{t}$ is straightforward. For rotation estimation, instead of estimating the periodic Euler angles, we represent the rotation target with a unit quaternion $\mathbf{q}$. It is a more friendly formulation as the network output.

Therefore, the output of our pose network is a vector including translation and unnormalized quaternion. We use the shared backbone as the encoder and add a decoder following [10]. Our baseline supervises the output with L1 loss:

$$\mathcal{L}_t = ||\mathbf{t} - \hat{\mathbf{t}}||_1, \quad \mathcal{L}_r = ||\mathbf{q} - \frac{\hat{\mathbf{q}}}{||\hat{\mathbf{q}}||}||_1, \quad \mathcal{L}_{pose} = \mathcal{L}_t + \lambda_r \mathcal{L}_r. \tag{7}$$

However, this loss design has several problems: 1) Adjusting the weight λ_r is difficult and expensive; 2) There is a domain gap for two 2D images to directly regress the 3D ego-motion; 3) We still need pose annotations during training. Therefore, we use a self-supervised loss [10,11] to replace it, considering its strength in these aspects. Specifically, the self-supervised loss is composed of an appearance matching loss $\mathcal{L}_p$ and a depth smoothness loss $\mathcal{L}_s$:

$$\mathcal{L}_{pose}(I_t, I_{t-\delta t}) = \mathcal{L}_p(I_t, I_{t-\delta t \to t}) + \lambda_s \mathcal{L}_s \tag{8}$$

Here $I_{t-\delta t \to t}$ represents the frame t synthesized with the image and depth of frame $t - \delta t$ and the predicted pose. More details are in the supplemental.

Note that in contrast to [10,11], we use the LiDAR signal to supervise the learning of depth directly and only use the self-supervised loss to learn pose. In this way, because the learning of depth is supervised by absolute depth values, we can also learn a metric-aware pose even without explicit pose annotations.

5 Experiments

5.1 Experimental Setup

Dataset We evaluate our method on the KITTI dataset [8]. It consists of 7481/7518 frames for training/testing and the training set is generally divided into 3712/3769 samples as training/validation splits. In this paper, apart from the multi-modality input data and annotations of the current frame, we also use three temporarily preceding frames. Related pose information is extracted from the raw data following Kinematic3D [3]. We use images and pose information of these preceding frames and only use LiDAR as depth supervision during training.
Metrics KITTI uses Average Precision (AP) for 3D object detection evaluation. It requires a 3D bounding box overlap of more than 70%/50%/50% for car/pedestrian/cyclist. We report the AP_{40} results following [32], corresponding to the AP of 40 recall points, which is more stable and fair for comparison.
Implementation Details We randomly select one of three temporarily preceding images together with the current frame as training input while use the earliest one during inference if not specified in experiments. Other hyper-parameter settings, data augmentation methods and loss designs basically follow recent binocular methods [5,12]. See more details in the supplemental materials.

Table 1. AP_{40} results on the KITTI validation benchmark.

Methods	Venue	AP_{3D} IoU$\geq$ 0.7			AP_{BEV} IoU$\geq$ 0.7		
		Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDIS [32]	ICCV 2019	11.06	7.60	6.37	18.45	12.58	10.66
MonoPair [6]	CVPR 2020	16.28	12.30	10.42	24.12	18.17	15.76
MoVi3D [31]	ECCV 2020	14.28	11.13	9.68	22.36	17.87	15.73
MonoDLE [23]	CVPR 2021	17.45	13.66	11.68	24.97	19.33	17.01
PGD [35]	CoRL 2021	19.27	13.23	10.65	26.60	18.23	15.00
CaDDN [26]	CVPR 2021	23.57	16.31	13.84	-	-	-
MonoFlex [44]	CVPR 2021	23.64	17.51	14.83	-	-	-
MonoRCNN [29]	ICCV 2021	16.61	13.19	10.65	25.29	19.22	15.30
GUPNet [21]	ICCV 2021	22.76	16.46	13.72	31.07	22.94	19.75
DFR-Net [49]	ICCV 2021	19.55	14.79	11.04	26.60	19.80	15.34
Kinematic3D [3]	ECCV 2020	19.76	14.10	10.47	27.83	19.72	15.10
DfM w/o pose	ECCV 2022	26.65	18.49	15.94	34.97	25.00	22.00
DfM w/ pose	ECCV 2022	**29.27**	**20.22**	**17.46**	**38.60**	**27.13**	**24.05**

5.2 Quantitative Analysis

Main Results First, we compare our framework with other state-of-the-art methods on the KITTI validation benchmark (Tab. 1), considering the ego-pose information is not available on the test set. We observe a significant improvement in both 3D detection and bird-eye-view (BEV) performance, 2.6%∼5.6% and 4.2%∼7.5% higher than the previous best for all the difficulty levels respectively. We conjecture that the better improvement on BEV performance is caused by our paradigm of voxel-based 3D detector: it finally detects 3D objects from the bird-eye-view following [42,16]. In addition, even without ego-pose information, our framework still outperforms others by a notable margin. This further shows the benefits brought by temporal information and stereo estimation. Please refer to the supplemental for its performance on the test set and other categories.

Comparison with Video-Based Methods Compared to the only previous methods using video information, Kinematic3D [3], our method also shows significant superiority. The reason is that Kinematic3D focuses more on the stability of detection and forecasting while our method pays more attention to depth estimation. Considering that the evaluation metric on KITTI requires particularly accurate localization for detected objects, our method naturally shows better performance on the benchmark. Note that our method is also compatible with some methods proposed in Kinematic3D. They can further improve the detection stability and efficiency of our framework and provide a natural integration with the downstream tasks such as tracking, prediction and planning.

Comparison with Binocular Methods Although our approach has achieved promising progress over previous monocular methods, we still observe a large gap between ours and binocular state of the art (64.7% AP for moderate). It is partly due to intrinsic weaknesses of the depth-from-motion setting. Nevertheless, we can expect a large space for improvement as the advancement of binocular methods, from RT3DStereo [14] (23.3% AP) to LIGA-Stereo [12] (64.7% AP).

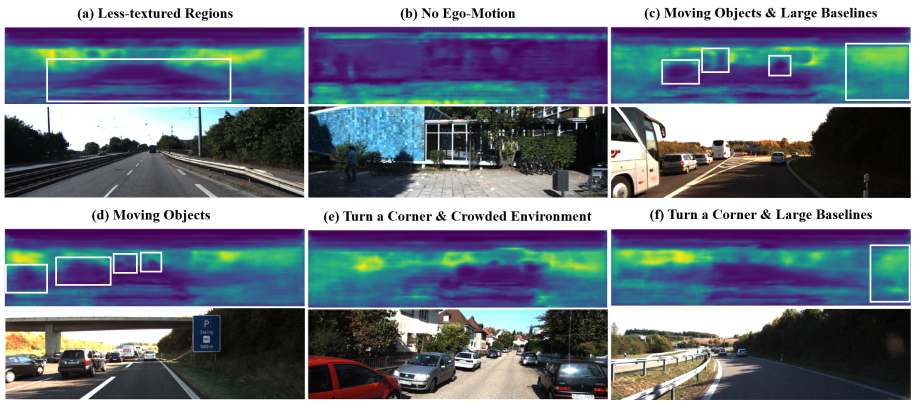

Fig. 5. Qualitative Analysis of aggregation weights in different cases. The depth estimation relies more on monocular priors for less textured regions in (a)(e), static cameras in (b) and moving objects in (c)(d) while tends to use stereo matching on other cases, especially on the background and the regions far away from camera centers in (f). Note that our analysis is still valid when the car is turning a corner in (e) because the rotation in the ego-motion is small in a short period.

5.3 Qualitative Analysis

For qualitative analysis, we show the visualization of aggregation weights (summed along the depth axis) in Sec. 4.3 with some representative cases (Fig. 5). For each sample plotted in the figure, we visualize the weight ranging from 0 to 1 above each image. Larger weights are marked with lighter regions in the weight maps, which indicates that the depth estimation relies more on stereo matching.

Next, we will discuss the inherent problems of stereo methods in the depth-from-motion setting analyzed in Sec. 3.3. In a general case, (a) shows that the estimation relies more on monocular priors for less textured regions such as the road. (b) shows a case that stereo matching will break down: no baseline is formed by static cameras. (c) and (d) show that stereo methods can not handle moving objects with the current pure design. In addition, on the right side of image (c), when the richness of texture seems similar, the regions far away from camera centers can form larger baselines. They can thus get more accurate estimations from stereo matching. A similar phenomenon can be seen in sample (f). Finally, even the driving car is turning a corner, all of our analysis is still valid because the rotation in the ego-motion can not be quite large in a short period. This weight is also learned adaptively for the crowded environment. These prove the interpretability of our method and the necessity of monocular compensation. It also points out possible directions for improving this group of the method, such as handling moving objects with customized designs in the stereo estimation.

For the visualization of 3D detection and depth estimation results from the perspective view and bird's eye view, please refer to the demo video attached in the supplementary material.

5.4 Ablation Studies

Geometry-Aware Stereo Cost Volume First, we show the benefits of geometry-aware stereo cost volume construction in Tab. 2. Both flip and rescale augmentation can remarkably enhance the detector. We suspect that making the cost volume more compatible with various augmented inputs can improve the generalization ability of models for different scenes and camera intrinsic settings.

Table 2. From top to down: Ablation studies for (a) geometry-aware cost volume construction, (b) detection performance of different depth estimation approaches, (c) using different preceding frames during inference, and (d) different pose-free designs.

Methods	AP_{3D} IoU$\geq$ 0.7			AP_{BEV} IoU$\geq$ 0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Baseline	17.41	12.93	11.60	24.78	18.21	16.06
+Flip aug.	19.13	13.92	12.62	26.89	19.48	17.52
+Rescale aug.	21.47	15.32	13.83	29.22	21.22	19.51
Mono Only	20.06	15.30	14.05	27.84	21.78	19.96
Stereo Only	21.47	15.32	13.83	29.22	21.22	19.51
Mono+Stereo	26.61	18.82	16.47	36.16	26.09	23.17
Prev-1st	24.09	17.27	15.03	35.50	25.24	22.82
Prev-2nd	24.92	17.62	15.68	35.89	25.39	22.99
Prev-3rd	25.19	17.96	15.92	36.16	25.88	23.03
Euler for rotation	20.16	15.03	13.01	28.96	21.21	19.08
+ quaternion	23.88	16.93	14.47	33.23	23.75	20.72
+ reproj. supervision	26.65	18.49	15.94	34.97	25.00	22.00

Monocular Compensation We compare different approaches for depth estimation in Tab. 2 and Fig. 6. We turn off one of two branches in Sec. 4.3 by setting the corresponding weight to zero during training and compare their detection (Tab. 2) and depth estimation accuracy (Fig. 6). We can see that with only monocular context, models still achieve a decent detection performance while failing on depth estimation of the entire scene. Stereo matching performs better on both aspects, especially the latter. Because these modules compensate each other fundamentally, our aggregation design brings an impressive gain thereon.

Different Preceding Frames As analyzed in Sec. 3.3, the distance of ego-vehicle in two frames can affect the baseline in this depth-from-motion setting and thus affect the accuracy of stereo matching. To compare the effect of using different frames, we train the model with a randomly selected previous frame for each sample and test it with a fixed one. Note that when the sample does not have the corresponding preceding frame, for instance, the third preceding one, we will use the earliest one that it has. As Tab. 2 shows, using the third preceding frame performs better than others up to about 1% mAP, which validates our analysis. This study has additional space for exploration: If given more previous frames, which one would be the best choice? If we involve multiple frames into stereo matching and depth estimation, what is a better frame selection design?

Pose-Free Designs Finally, we study the specific designs for pose-free depth from motion. Our baseline uses the Euler angle as the rotation representation as [3] and directly regresses the translation and rotation with the pose supervision. We further try the quaternion representation and reprojected photometric loss as the supervision, and both show superiority than before. More importantly, we can avoid the pose annotation completely with the self-supervised paradigm, which is especially important for the practice in the real world.

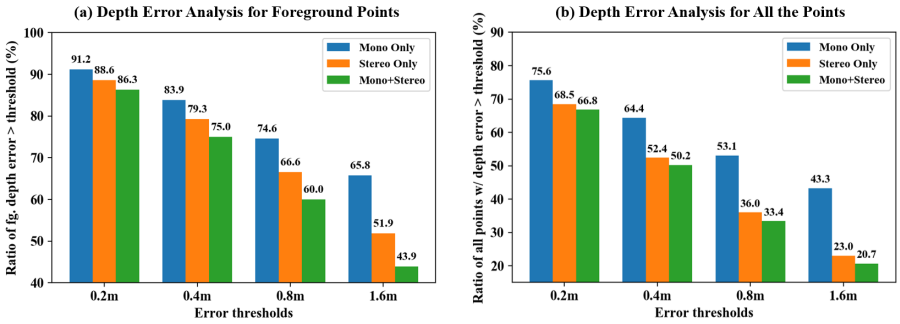

Fig. 6. We make error analysis for the depth predictions of foreground region and the entire scene by different methods, respectively, by comparing the percentage of points with depth errors greater than thresholds: 0.2m, 0.4m, 0.8m, 1.6m. The error medians of monocular/stereo/hybrid methods on the foreground region/the entire scene are 5.86/3.33/2.60m and 1.15/0.58/0.48m.

6 Conclusion

In this paper, we propose a framework for monocular 3D detection from videos. It lifts 2D image features to 3D space via an effective depth estimation module and detects 3D objects on top. The depth-from-motion system leverages an important ego-motion clue to estimate depth from stereo matching, which is further compensated with monocular understanding for addressing several intrinsic dilemmas. To make this framework more flexible, we further extend it to pose-free case with an effective rotation formulation and a self-supervised paradigm. Experimental results show the efficacy of our method and validate our theoretical discussion. In the future, we will optimize our framework in terms its simplicity and generalization ability. How to address the stereo estimation of moving objects is also an important problem worthy of further exploration.

Acknowledgements This work is supported by GRF 14205719, TRS T41-603/20-R, Centre for Perceptual and Interactive Intelligence, and CUHK Interdisciplinary AI Research Institute.

References

1. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 331–346 (2018)
2. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: IEEE International Conference on Computer Vision (2019)
3. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: Proceedings of the European Conference on Computer Vision (2020)
4. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Conference on Neural Information Processing Systems (2015)
5. Chen, Y., Liu, S., Shen, X., Jia, J.: Dsgn: Deep stereo geometry network for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12536–12545 (2020)
6. Chen, Y., Tai, L., Sun, K., Li, M.: Monopair: Monocular 3d object detection using pairwise spatial relationships. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
7. Garg, R., BG, V.K., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: Proceedings of the European Conference on Computer Vision (2016)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
9. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with leftright consistency. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
10. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
11. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494 (2020)
12. Guo, X., Shi, S., Wang, X., Li, H.: Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3153–3163 (2021)
13. Jörgensen, E., Zach, C., Kahl, F.: Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. CoRR **abs/1906.08070** (2019), https://arxiv.org/abs/1906.08070
14. Königshof, H., Salscheider, N.O., Stiller, C.: Realtime 3d object detection for automated driving using stereo vision and semantic information. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 1405–1410. IEEE (2019)
15. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1611–1621 (2021)
16. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)

17. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7644–7652 (2019)
18. Li, P., Zhao, H., Liu, P., Cao, F.: Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In: European Conference on Computer Vision (2020)
19. Liu, C., Gu, J., Kim, K., Narasimhan, S.G., Kautz, J.: Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10986–10995 (2019)
20. Liu, M., Zhu, M.: Mobile video object detection with temporally-aware feature maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5686–5695 (2018)
21. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
22. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (ToG) **39**(4), 71–1 (2020)
23. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
24. Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
25. Peng, X., Zhu, X., Wang, T., Ma, Y.: Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 119–128 (2022)
26. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distributionnetwork for monocular 3d object detection. CVPR (2021)
27. Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. CoRR **abs/1811.08188** (2018), https://arxiv.org/abs/1811.08188
28. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
29. Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.K.: Geometry-based distance decomposition for monocular 3d object detection. In: IEEE International Conference on Computer Vision (2021)
30. Simonelli, A., Bulo, S.R., Porzi, L., Kontschieder, P., Ricci, E.: Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3225–3233 (2021)
31. Simonelli, A., Bulò, S.R., Porzi, L., Ricci, E., Kontschieder, P.: Towards generalization across depth for monocular 3d object detection. In: Proceedings of the European Conference on Computer Vision (2020)
32. Simonelli, A., Bulò, S.R.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: IEEE International Conference on Computer Vision (2019)
33. Sun, J., Chen, L., Xie, Y., Zhang, S., Jiang, Q., Zhou, X., Bao, H.: Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10548–10557 (2020)

34. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. arXiv preprint arXiv:1812.04605 (2018)
35. Wang, T., Xinge, Z., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: Conference on Robot Learning. pp. 1475–1485. PMLR (2022)
36. Wang, T., Zhu, X., Lin, D.: Reconfigurable voxels: A new representation for lidar-based point clouds. In: Conference on Robot Learning (2020)
37. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2021)
38. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
39. Xiao, F., Lee, Y.J.: Video object detection with an aligned spatial-temporal memory. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 485–501 (2018)
40. Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
41. Xu, Z., Zhang, W., Ye, X., Tan, X., Yang, W., Wen, S., Ding, E., Meng, A., Huang, L.: Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12557–12564 (2020)
42. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10) (2018)
43. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5336–5345 (2020)
44. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
45. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
46. Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., Lin, D.: Ssn: Shape signature networks for multi-class object detection from point clouds. In: Proceedings of the European Conference on Computer Vision (2020)
47. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7210–7218 (2018)
48. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 408–417 (2017)
49. Zou, Z., Ye, X., Du, L., Cheng, X., Tan, X., Zhang, L., Feng, J., Xue, X., Ding, E.: The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)