ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095283

OVERCOMING THE SEESAW IN MONOCULAR 3D OBJECT DETECTION VIA LANGUAGE KNOWLEDGE TRANSFERRING

Weichen Xu † *Tianhao Fu* †

Peking University, Beijing, China

ABSTRACT

Monocular 3D object detection is a challenging problem in self-driving and computer vision communities. Previous works suffered from a severe seesaw phenomenon: multi-category learning was worse than single-category, and feature learning between categories inhibited each other. We reveal that the real culprit is the significant difference in depth distribution between categories. Confusing feature representations exacerbate depth estimation. In this paper, we propose Language Knowledge Transferring to introduce language information in monocular 3D object detection, termed as MonoLT. Multimodal language-Image guides networks learn more class-specific features, which reduces the pressure of depth estimation. Meanwhile, we propose the Polar Depth Aggregator to make the depth estimation less disturbed by the environment and other instances (especially different classes). Comprehensive experiments performed on the KITTI dataset prove the superiority of our proposed method. The code will be released soon.

***Index Terms*—** Monocular 3D Detection, Multimodal language-Image learning, Monocular depth estimation

1. INTRODUCTION

Monocular 3D object detection is a challenging task. Its goal is to estimate the 3D boxes in a single image [1], including 3D coordinates, 3D dimensions, and rotation angles. Due to its low price and configuration simplicity, it can be applied to self-driving and sweeping robots. Some methods [2, 3] leverage monocular depth networks to generate pseudo point clouds. Then superior point cloud-based 3D detection networks are utilized. However, the additional depth estimator incurs significant overhead in inference. Another line of work [4, 5] directly regress 3D parameterizations from standard 2D detectors [6, 7]. The simple end-to-end architecture is capable of learning geometry-aware representation and achieving competitive detection accuracy, which is worthy of research.

Some end-to-end monocular 3D object detection approaches have achieved impressive performance. However, state-of-the-art (SOTA) approaches still suffer from a severe seesaw phenomenon in autonomous driving datasets, as shown in Fig. 1(a). We use DID-M3D [5] as baseline and find that training with only the Car category performs significantly better than training with all three categories, Pedestrian, Car, and Cyclist, indicating that feature learning in different categories inhibits each other. To find out the cause of this embarrassing problem, on the one hand, we take KITTI [8] as an example to visualize the depth distribution of the above three categories, i.e., the range of depth and the depth unsmoothness. Depth unsmoothness is defined as the absolute difference between adjacent regions when the instance is quadratically divided, and the result is shown in Fig. 1(b). It can be seen that there are apparent differences between the three categories. Cars usually appear at a distance, while pedestrians and cyclists are opposite (possibly because the two categories are too small to be labeled at a distance). On the other hand, we use a well-trained DID-M3D object detector to extract the features from the ground truth before depth estimation, as shown in Fig. 1(c). Principal component

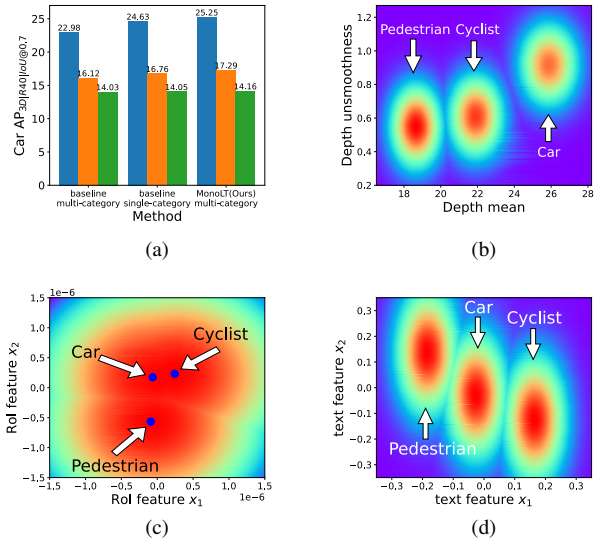

Fig. 1. (a) The seesaw phenomenon in the autonomous driving dataset. The three columns in each method correspond to easy, moderate, and hard. (b) The depth distribution of Pedestrian, Car, and Cyclist. (c) Visualization of features extracted from a well-trained DID-M3D object detector. (d) The text features of the above three categories, generated from the CLIP text encoder.

† Equal contribution.

analysis (PCA) [9] is utilized to reduce dimensions for visualization. It can be seen that even on the scale of 10^{-6}, the features of the three categories are still mixed. When we train a single-category detector, such as Car, the network already implicitly contains a priori for that class, and the regression network only needs to estimate depth in the range of the Car to obtain higher accuracy. On the contrary, when the three categories are trained at the same time, the confusing feature representations make it difficult for the network to learn the mapping with depth. Therefore, in monocular 3D object detection, it is necessary to learn more class-specific features to reduce the pressure of depth estimation.

Learning from multimodal data has become popular because vision tasks are assisted by language, audio, etc., and have shown promising results. Contrastive Language-Image Pre-Training (CLIP) [10] trains a huge image-text dataset in a contrastive learning fashion. As a result, paired image encoders and text encoders can learn interactive feature representation. The barriers between the two modalities are bridged, which helps both vision and language tasks. In addition to impressive success in image classification [11], the same achievements have been made in image segmentation [12, 13] and object detection [14, 15]. Motivated by the outstanding performance of language-Image learning on visual representation learning, we use a well-trained CLIP text encoder [10] to generate text features of {Pedestrian}, {Car}, and {Cyclist}. Specifically, we feed the category names into prompt templates and use an ensemble of various prompts. For example, 'a photo of one {category}' is constructed and used for feature extraction. Following [10], a list of 63 prompt templates is used for each category. Then, PCA is used to reduce dimensions, and the results are shown in Fig. 1(d). Surprisingly, there is a relatively clear distinction between the text features of the three categories. Inspired by this, we propose Language Knowledge Transferring. After obtaining the region of interest (RoI) features, we use the fully connected layer to compute logits whose weights are initialized by the integrated text features of three categories generated by the CLIP text encoder. Through end-to-end training, the discrimination of text features is transferred into visual features. With multimodal language-Image learning, the network can extract more class-specific features.

Although more class-specific features alleviate the pressure on depth estimation, it is essential to ensure that background and other instances (especially different classes) are excluded. Accordingly, we propose the Polar Depth Aggregator (PDA) to adaptively improve the attention between regions within the instance and exclude other interferences. To adapt to the individual shapes of the three categories, we add supervision to the offset of deformable convolution [16] to focus on nine polar boundaries within the instance. Compared to learning offsets randomly, purposeful supervision can avoid interference from external regions. As a result, self-information is fused while irrelevant information is excluded, which complements Language Knowledge Transferring and is critical for information-sensitive tasks like depth estimation.

To sum up, our main contributions are threefold:

- We propose Language Knowledge Transferring to learn more class-specific features. We are the first to introduce language knowledge in monocular 3D object detection.
- We propose Polar Depth Aggregator (PDA) to fuse self-information and exclude the interference of irrelevant regions for depth estimation.
- Comprehensive experiments performed on the KITTI dataset prove the superiority of our proposed method. Our approach achieves state-of-the-art results by introducing language knowledge in monocular 3D object detection.

2. METHODOLOGY

2.1. Overall Pipeline

The MonoLT we proposed is based on remarkable end-to-end monocular 3D object detection methods such as DID-M3D [5]. Fig. 2 is an overview of the proposed method, which shows that our Language Knowledge Transferring is parallel to other 3D branches, and the Polar Depth Aggregator is inserted before depth estimation.

2.2. Language Knowledge Transferring

After obtaining the RoI features $\mathbf{f}_r$, we add a branch in parallel for introducing language knowledge. Specifically, a series of convolutions are utilized to extract features further. Finally, we use the fully connected layer to compute logits. Fig. 2 shows the architecture and training objective.

On the other hand, we generate the text embeddings by feeding the category texts {Pedestrian}, {Car} and {Cyclist} with prompt templates, e.g., "a photo of one {category}", into the pretrained CLIP text encoder [10]. Multiple prompt templates are ensembled into the final text features $\mathbf{t}_{\text{Pedestrian}}$, $\mathbf{t}_{\text{Car}}$, and $\mathbf{t}_{\text{Cyclist}}$.

The weights of the last fully connected layer are initialized by the text features. In this way, similarity can be calculated between region features and all text embeddings:

$$\mathbf{z}(r) = [sim(\mathbf{f}_r, \mathbf{t}_{\text{Pedes}}), sim(\mathbf{f}_r, \mathbf{t}_{\text{Car}}), sim(\mathbf{f}_r, \mathbf{t}_{\text{Cyc}})]. \quad (1)$$

Then we apply softmax activation with a temperature τ to compute the cross entropy loss:

$$\mathcal{L}_{language} = \mathcal{L}_{CE}(softmax(\frac{\mathbf{z}(r)}{\tau}), y_r), \quad (2)$$

where y_r denotes the class label.

Language Knowledge Transferring is used only in training to transfer the discrimination of text features to visual features, which will not increase any cost during inference.

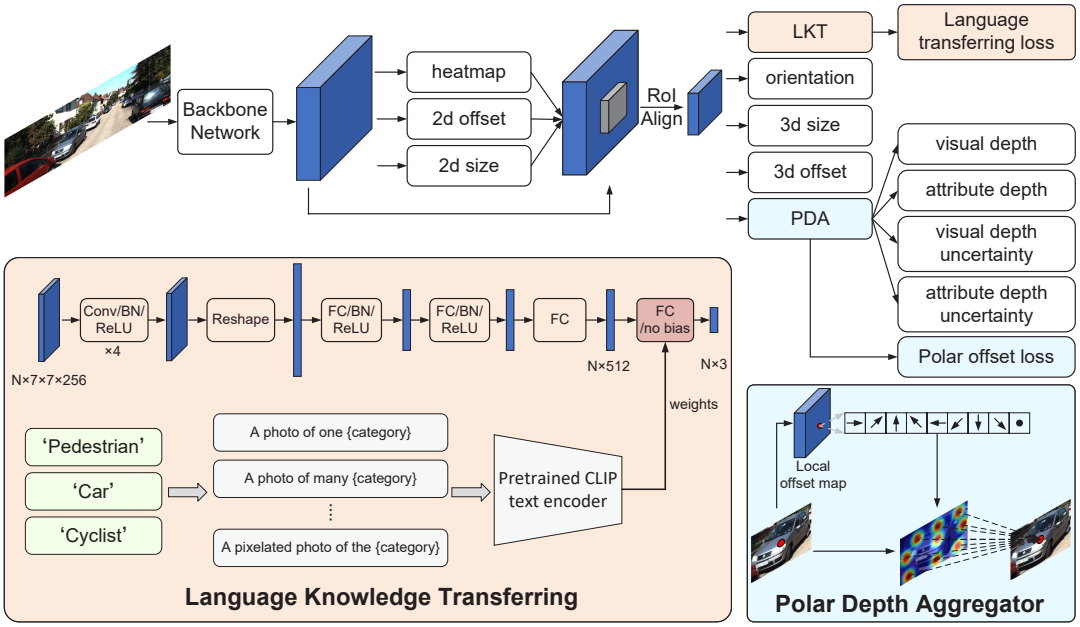

Fig. 2. Architecture overview of MonoLT. Our proposed method contains two components: Language Knowledge Transferring and Polar Depth Aggregator. The dark blue cubes represent the features, and the unfilled rectangles represent the corresponding networks. In Polar Depth Aggregator, only one pixel inside the car is visualized, and the same at the remaining pixels.

2.3. Polar Depth Aggregator

There is no supervision for offsets in standard deformable convolution [16] and plain convolution. Optimizing the output features will introduce interference from irrelevant regions, which exacerbates the depth estimation. To this end, we propose a Polar Depth Aggregator (PDA) module inserted before the depth estimation. The offset in deformable convolution is manually guided to nine polar boundaries of the instance. The illustration of our PDA module is shown in Fig. 2. Local offset map is predicted from the RoI feature by the convolution layer as follows:

$$\triangle P_j = \mathrm{Conv}(\mathbf{f}_r(p_j)), \tag{3}$$

where p_j is the location of a pixel, $\mathbf{f}_r(p_j)$ is the RoI feature of the j-th pixel, and $\triangle P_j = \left\{\triangle p_j^o | o = 1, \ldots, 9\right\}$ is the local offset map. This will be used in deformable convolution to extract non-grid-like features. As suggested in [16], the sum of pixels p_j and offsets $\triangle P_j$ is used as the attention position. All the features of the attention position are fused as a new feature:

$$\mathbf{f}_r^{'}(p_j) = \sum_{o=1}^{9} w_o \cdot \mathbf{f}_r(p_j + \triangle p_j^o). \tag{4}$$

To facilitate the later depth estimation, we impose additional supervision for the offsets $\triangle P_j$ and compute an auxiliary loss to guide the offsets. Eight positions at the polar boundaries are obtained at 45° intervals for each pixel in each instance. Combined with its own position, the ground truth coordinates $G_j = \left\{g_j^o | o = 1, \ldots, 9\right\}$ corresponding to the nine offsets is obtained. Then, we can calculate the ground truth of offsets by $\triangle g_j^o = g_j^o - p_j$.

No special matching, for example, the Hungarian algorithm, is required. The nine polar directions are just assigned to nine offsets of deformable convolution. We use Smooth L1 loss to supervise the offset between $\triangle P_j$ and $\triangle G_j$:

$$\mathcal{L}_{offset} = \frac{1}{9N_p} \sum_{j=1}^{N_p} \sum_{o=1}^{9} SmoothL1(\triangle p_j^o, \triangle g_j^o), \tag{5}$$

where N_p denotes the number of pixels inside an instance.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experimental Setup

We use the KITTI [8] dataset to report the performance of our method. The KITTI dataset contains 7481 training images and 7518 test images. Following the suggestion of [1], we divided the training images into a training set (3712) and a val set (3769). The results of the val set are reported on the training set (3712), while the results of the test set are reported on all 7481 training images.

Following [5], the size of the input image is zero-padded to 384×1280. Random horizontal flip is the only data augmentation. We use PyTorch and a total batch size of 16 on 2 NVIDIA TITAN Xp GPUs (8 images per GPU). We train the

Fig. 3. Some qualitative 3D detection results on the KITTI test set.

Table 1. Comparison with other state-of-the-art methods on KITTI test set. The IoU thresholds for AP_{3D} are 0.7 for Car.

method	$AP_{3D}\|R_{40}$			$AP_{BEV}\|R_{40}$		
	Easy	Mod	Hard	Easy	Mod	Hard
SMOKE [17]	14.03	9.76	7.84	20.83	14.49	12.75
PGD [4]	19.05	11.76	9.39	26.89	16.51	13.49
CaDDN [18]	19.17	13.41	11.46	27.94	18.91	17.19
MonoFlex [19]	19.94	13.89	12.07	28.23	19.75	16.89
DCD [20]	23.81	15.90	13.21	32.55	21.50	18.25
MonoCon [21]	22.50	16.46	13.95	31.12	22.10	19.00
DID-M3D [5]	24.40	16.29	13.75	32.95	22.76	19.83
MonoLT	**25.51**	**17.18**	**14.56**	**34.35**	**23.82**	**20.80**
Improvements	+1.11	+0.89	+0.81	+1.40	+1.06	+0.97

network with 150 epochs. AdamW optimizer is adopted, and the initial learning rate is set to 0.001.

3.2. Comparison to State-of-the-Art

We compare our approach with other works that report state-of-the-art performance on the KITTI [8] test set, as shown in Table 1. Our method achieves superior performance on both AP_{3D} and AP_{BEV}. We outperform our baseline DID-M3D [5] by 1.11 in AP_{3D} Easy level and 0.89 in Mod. level, which reveals that language knowledge is critical to learning class-specific features.

3.3. Ablation Study

In Fig. 1(a), we perform multi-category learning in MonoLT and compare it with baseline under single-category and multi-category. Our approach significantly outperforms the baseline under multi-category and slightly outperforms the baseline under single-category, which indicates that multimodal language-Image learning can guide the network to learn more class-specific features and alleviates the seesaw phenomenon caused by multi-category feature learning.

Table 2. Ablation study on KITTI val set.

LKT	PDA	$AP_{3D}\|R_{40}$		
		Easy	Moderate	Hard
		22.98	16.12	14.03
✓		24.51	16.82	13.96
	✓	23.52	16.23	13.46
✓	✓	**25.25**	**17.29**	**14.16**

There are two components in our MonoLT, Language Knowledge Transferring (LKT), and PDA. As shown in Table 2, both LKT and PDA play significant roles in our approach. LKT can achieve an improvement from 22.98% AP to 24.51% AP in the Easy level and achieve 0.7 points improvement in the Mod. level, which has been aligned with the baseline under single-category. PDA can exclude the interference during depth estimation and can achieve an improvement from 22.98% AP to 23.52% AP in the Easy level.

Table 3. Influence of language knowledge compared to hard-coding on KITTI val set.

method	$AP_{3D}\|R_{40}$		
	Easy	Moderate	Hard
baseline	22.98	16.12	**14.03**
baseline + hard-coding	23.95	16.60	13.73
baseline + LKT	**24.51**	**16.82**	13.96

In order to prove the superiority of multi-modal language-Image learning, we hard-coded the classification results obtained by heatmap after RoI features, as shown in Table 3. The hard-coding is a explicit way to introduce class-specific features. Specifically, ROI Align extracts $7 \times 7 \times 64$ features, and then directly concatenate the categorical result of heatmap $7 \times 7 \times 3$ in the first stage, and outputs $7 \times 7 \times 67$. It can be seen that hard-coding can make features more class-specific but worse than LKT. Our LKT implicitly guides the network to learn language knowledge during training without increasing inference time and memory cost.

3.4. Visualization

In Fig. 3, we provide the qualitative results of the proposed method on the KITTI test set. The proposed method predicts the location, size, and orientation well in the test set.

4. CONCLUSIONS

In this work, we point out that the seesaw phenomenon is caused by the fact that confusing feature representations exacerbate the depth estimation. To alleviate this problem, we have proposed MonoLT, a novel monocular 3D object detection method supported by language knowledge. Multimodal language-Image learning can guide the network to learn more class-specific features and alleviates the pressure of depth estimation. Moreover, we have proposed the Polar Depth Aggregator to exclude the interference of irrelevant information. Equipped with these components, MonoLT beats existing state-of-the-art approaches on KITTI, which shows the superiority of our method.

5. REFERENCES

[1] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "3d object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems*, 2015.

[2] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang, "Rethinking pseudo-lidar representation," in *European Conference on Computer Vision*. August 2020, pp. 311–327, Springer.

[4] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Proceedings of the 5th Conference on Robot Learning*, 2022.

[5] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai, "Did-m3d: Decoupling instance depth for monocular 3d object detection," in *European Conference on Computer Vision*, 2022.

[6] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] Wold Svante, Esbensen Kim, and Geladi Paul, "Principal component analysis," in *Chemometrics and Intelligent Laboratory Systems*, August 1987, pp. 37–52.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *arXiv preprint*, vol. abs/2109.01134, 2021.

[12] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2022.

[13] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18134–18144.

[14] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *European Conference on Computer Vision*, 2022.

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10965–10975.

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[17] Zechen Liu, Zizhang Wu, and Roland Toth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[18] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8555–8564.

[19] Yunpeng Zhang, Jiwen Lu, and Jie Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang, "Densely constrained depth estimator for monocular 3d object detection," in *European Conference on Computer Vision*, 2022.

[21] Xianpeng Liu, Nan Xue, and Tianfu Wu, "Learning auxiliary monocular contexts helps monocular 3d object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.