AEAM3D:ADVERSE ENVIRONMENT-ADAPTIVE MONOCULAR 3D OBJECT DETECTION VIA FEATURE EXTRACTION REGULARIZATION

Yixin Lei[1], *Xingyuan Li*[1], *Zhiying Jiang*[1], *Xinrui Ju*[1], *Jinyuan Liu*[1*]

[1] Dalian University of Technology, China

yixinlei@mail.dlut.edu.cn, atlantis918@hotmail.com

ABSTRACT

3D object detection plays a crucial role in intelligent vision systems. Detection in the open world inevitably encounters various adverse scenes while most of existing methods fail in these scenes. To address this issue, this paper proposes a monocular 3D detection model, termed AEAM3D, which effectively mitigates the degradation of detection performance in various harsh environments. Additionally, we assemble a new adverse 3D object detection dataset encompassing some challenging scenes, including rainy, foggy, and low light weather conditions. Experimental results demonstrate that our proposed method outperforms current state-of-the-art approaches by an average of 3.12% in terms of AP_{R40} for car category across adverse environments.

***Index Terms*—** 3D object detection, monocular vision, image enhancement

1. INTRODUCTION

Generally vision-based object detection plays a critical role in autonomous driving while 3D object detection remains a complex task. A common approach frequently utilize LiDAR sensors or stereo cameras for depth estimation[1, 2, 3]. Yet they significantly increase the cost of implementing practical systems [4]. Consequently, monocular 3D object detection[5, 6, 7] has emerged as a promising alternative.

Existing monocular 3D object detection techniques can be broadly classified into two categories: those based on single images, such as M3D-RPN[8] and MonoDLE[5] and those leveraging auxiliary information, including RoI-10D[9] and Pseudo-LiDAR[10]. Nevertheless, these methods have encountered several issues. (i) 3D object detection is inevitable to face real-world adverse conditions, causing degraded image quality [11, 12]. (ii) Monocular 3D detection is inherently limited by the single viewpoint, leading to uncertainties in depth estimation. (iii) The scarcity of datasets induces the failure of capturing adverse weather conditions characteristic.

(a)GUPNet (b)Retorer+3D Detection (c)AEAM3D(Ours)

Fig. 1. The comparison illustrating 3D object detection methods in adverse scenes: (a) existing models overlooking environmental context; (b) conventional solutions using image restoration, potentially yielding unsuitable images; (c) AEAM3D employing adaptive learning strategy which penalizes perceptual errors.

To address these issues, we propose a novel monocular 3D object detection method, dubbed AEAM3D. Our approach incorporates an adaptive learning strategy allowing better adaption to complex scenarios. In addition, we introduce a diverse dataset, including seven harsh conditions. Figure 1 demonstrates that our proposed model outperforms state-of-the-art(SOTA) 3D object detectors and cascade of image enhancement and 3D detection models. Our contributions are three-fold:

- We introduce a robust network specifically designed to handle a variety of adverse environments, significantly improving the performance of monocular 3D object detection models across a wide range of challenging real-world situations.
- We propose an adaptive learning strategy during the training process to extract resilient features that remain less susceptible to degrading factors, aiding the model in discerning various inclement environments.
- To support 3D object detection in harsh environments, we have compiled a comprehensive dataset comprising 7,481 images for seven demanding conditions.

*Corresponding author.

This work was partially supported by China Postdoctoral Science Foundation (2023M730741), and the National NaturalScience Foundation of China (No.62302078).

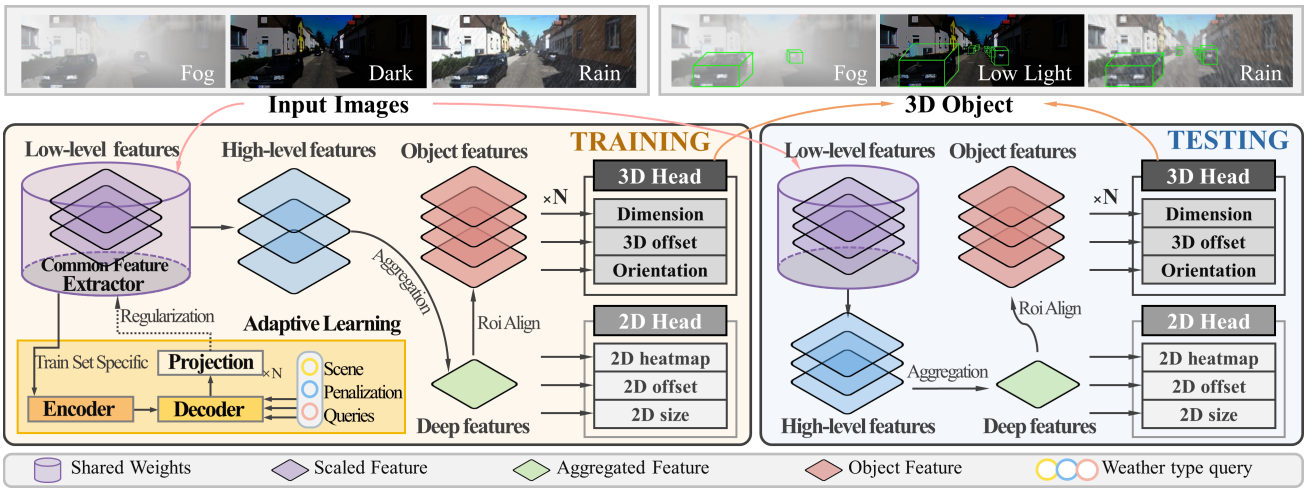


Fig. 2. The pipeline of AEAM3D. The core components, adaptive learning (Sec. 3.2) regularizes features from Common Feature Extractor to help model perceive clean meta features that are not degraded by adverse factors.

2. METHOD

Under inclement weathers, the specific spectral interaction between objects and camera may be affected by absorption and scattering of suspended water droplets, dust, and other particulates, inducing the loss of depth information [13]. Thus we present a monocular 3D object detection model for adverse environments, namely AEAM3D.

2.1. Adverse Condition Datasets Generation

We compile datasets encompassing several adverse weather conditions. The process of synthesizing these conditions bases on their corresponding atmospheric effects. The fog condition is modeled based on the atmospheric light attenuation theory [14], and rain with rain streaks and fog effect is on [15]. In addition, the image brightness is reduced using luminance correction method to simulate low light conditions.

2.2. Adaptive Learning Strategy

We propose a novel adaptive learning strategy comprising an encoder and a decoder, as depicted in Figure 2, which are specifically designed to act as a constraint, rather than focus on image restoration. Particularly the encoder assists the model in rectifying inaccurate feature perception under adverse conditions. The decoder employs learnable scene penalization queries to penalize incorrect perception by which the model can suppress potential errors. Notably, this learning strategy is only required during training.

Given a degraded image I of size $H \times W \times 3$, a common feature extractor is applied to generate low-level features $(\frac{H}{4} \times \frac{W}{4} \times C)$. These features are then put into the encoder which employs SwinBlocks at different stages.

Encoder: During each stage we use patch merging where the resolution is reduced to assist the module in learning both coarse and fine contents, and the merged features are passed on to the subsequent stage. SwinBlocks are then perform feature transformation while maintaining the resolution. A SwinBlock comprises a shifted window-based $\mathrm{MSA}_{\mathrm{SW}}$ and an MLP. Layer Normalization ((LN) is applied prior to each $\mathrm{MSA}_{\mathrm{W}}$ and MLP module, and a residual connection is incorporated each module. $\mathrm{MSA}_{\mathrm{W}}$ and $\mathrm{MSA}_{\mathrm{SW}}$ denote window based self-attention using traditional and shifted window. Following [16], self-attention is calculated as:

$$\mathbf{Attention}(Q, K, V) = \mathbf{SoftMax}\left(QK^T/\sqrt{d} + B\right)V, \quad (1)$$

where Q, K, V are queries keys and values that have same dimensions. B is relative position bias.

Decoder: In the decoder, scene penalization queries are utilized to output a task feature vector. Cross-attention is applied in this module, with K and V taken from the same output features as the last stage of the encoder, and Q being the learnable queries. The output features of the decoder serve as the weather type task vector and are fused with the features produced by each stage of the encoder.

2.3. 3D Object Detection in Adverse Scenes

Figure 2 shows the framework of our approach, monocular 3D object detection takes an RGB image as input and constructs a 3D bounding box for the object in 3D space. Concretely, 2D detection backbone from low-level constraint features is applied to produce high-level deep features, and then these features are aggregated to get deep features with resolution $\mathbf{F} \in \mathbf{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$. Subsequently, we apply three 2D detection heads in deep features $\mathbf{F}$ to predict 2D heatmap $\mathbf{H}$. Through

using ROI_{Align} in deep feature map with 2D box information, the features are generated whose size is 7×7 and finally used in the 3D detection heads to predict the object 3D center offset O_{3d}, 3D size S_{3d} and direction Θ.

2.4. Loss Functions

Throughout the training process, we concurrently compute the losses associated with the adaptive learning strategy and the 3D object detection task. Our adaptive learning strategy employs the $smooth_{L1}$ loss to penalize the incorrect perception of features in challenging scenarios. This loss function is formulated as follows:

$$\mathcal{L}_{smooth\,L_1} = \begin{cases} 0.5\mathbf{E}^2 & \text{if } |\mathbf{E}| < 1 \\ |\mathbf{E}| - 0.5 & \text{otherwise}, \end{cases} \quad (2)$$

where $\mathbf{E}$ represents the difference between the perceived scene and real scene.

For 3D object detection, the loss function is as the following formula, containing 2D detection part and 3D detection part. The 2D offset O_{2D} refers to the residual towards rough 2D centers and $\mathcal{S}_{2D}$ denotes the 2D box height and width. We follow [17] to use loss functions $\mathcal{L}_H$, $\mathcal{L}_{O_{2d}}$, $\mathcal{L}_{S_{2d}}$. For the dimensions of the 3D object, we use the typically designed $\mathcal{L}_{S_{3d}}$ and multi-bin to calculate $\mathcal{L}_\Theta$ for the prediction of the object observation angle. The instance depth loss is $\mathcal{L}_{D_{ins}}$. We set the weight of each loss term to 1.0. The overall loss is:

$$\mathcal{L} = \mathcal{L}_H + \mathcal{L}_{O_{2d}} + \mathcal{L}_{S_{2d}} + \mathcal{L}_{S_{3d}} + \mathcal{L}_\Theta + \mathcal{L}_{O_{3d}} + \mathcal{L}_{D_{ins}}. \quad (3)$$

3. EXPERIMENTS

This section compares the results of our method for 3D object detection in various adverse environments. Our experiments are performed on 4 Nvidia TITAN XP and a batch size of 8.

3.1. Datasets and Metric

We evaluate our methods and 5 state-of-the-art under synthetic KITTI 3D dataset. Following the methodology of [18], the dataset is partitioned into 3,712 sub-training sets and 3,769 validation sets. Detection outcomes are presented in three levels of difficulty, namely easy, moderate, and hard. We use average precision as the evaluation metric. We train the network for 140 epochs, following the Hierarchical Task Learning (HTL) strategy. Input images are resized to a resolution of 1280 × 384, with pixel values in the range of [0, 255]. The pixel intensities are then adjusted based on the mean pixel intensity of the entire dataset.

3.2. Comparison with 3D Detection Methods

This section conducts a comprehensive comparison between AEAM3D and several SOTA monocular 3D object detection techniques. The car category's 3D detection accuracy, denoted by $\mathrm{AP3D}_{R40}$, serves as the benchmark for comparison.

As presented in Table 1, our method achieves significant performance improvements across different weather conditions. Under the heavy rain dataset, our method exceeds DID-3D by 1.66%, 0.98%, and 0.87% on easy, moderate and hard settings. On the thick fog, our method outperforms GUPNet by 3.33%, 2.04%, and 1.78% at a 0.7 IoU threshold. Meanwhile AEAM3D substantially surpasses DID-M3D and MonoDLE in the low light, with improvements of 0.91% and 3.71% $\mathrm{AP3D}_{R40}$ under moderate setting.

3.3. Comparison with Restoration Methods

In this section, we extend to compare our base 3D detection network with various image restoration techniques, as shown in the table 2. TransWeather is trained in these three weather conditions since it is designed to adapt to various weather scenarios. Other methods are trained under specific environments tailored to their corresponding effects.

The performance of AEAM3D is comparable to TransWeather in dense fog, but shows significant improvement under heavy rain conditions. Under low light conditions it has significantly improved by 7.85%, 5.17%, and 3.64% under the three settings of easy, mod, and hard, respectively. In addition, our method is also significantly superior to all other task specific methods. For example, under dense fog conditions, our method improved by 3.13%, 2.14%, and 1.80% compared to MSBDN under three different settings, respectively. In summary, our proposed method not only attains state-of-the-art accuracy in harsh environments when compared to the leading 3D object detection techniques, but it also outperforms existing image restoration methods.

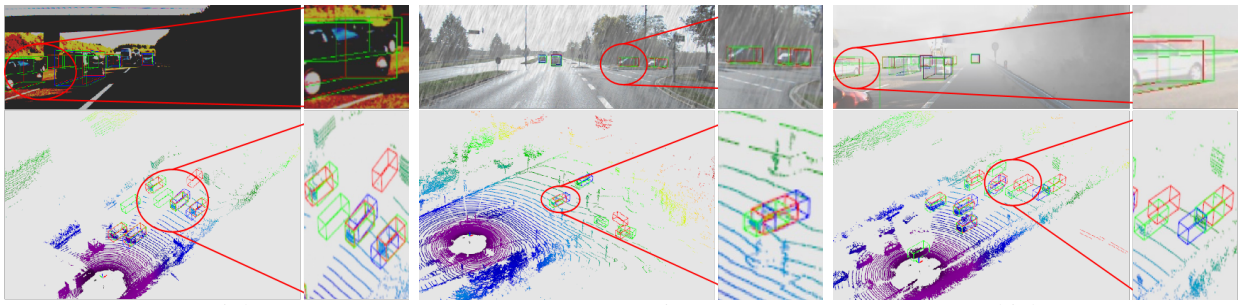
3.4. Ablation Study

To investigate how each module in AEAM3D enhances detection, we randomly selected one seventh of medium rain, heavy rain, dense rain, thin fog, thick fog, dense fog, and low light to obtain a mixed dataset, and then tested each module on this dataset. The results are shown in Table 3.

We evaluate the effectiveness of our adaptive learning strategy by examing respectively the impact of the encoder and the decoder on the overall performance. The encoder's contribution is examined by comparing settings (a→b). The results demonstrate that the encoder consistently improves the overall performance by 0.34% for (a→b) under moderate settings. Then we assess the decoder through experiments (a→c). At the same time the improvements of experiments (b→d) and (c→d) indicate that both parts of the module prove to be indispensable for optimal performance.

Table 1. Comparison of the latest 3D object detection methods on the moderate fog, thick fog, moderate rain, heavy rain, dense rain and low light dataset based on AP_{3D} of car category.

Methods	Venue	Mod. Fog			Thick Fog			Mod. Rain			Heavy Rain			Dense Rain			Low Light		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
SMOKE	*CVPR20*	8.86	5.98	4.53	5.10	3.31	2.28	7.33	5.24	4.03	5.97	3.78	2.77	5.64	3.88	3.21	5.48	4.03	3.49
MonoFLEX	*CVPR21*	19.97	14.11	11.86	18.37	13.28	10.57	17.21	12.94	11.55	16.99	11.83	10.12	15.35	12.14	10.38	10.43	8.32	7.75
MonoDLE	*CVPR21*	14.77	12.15	10.02	17.35	12.89	11.27	15.65	13.34	12.33	15.64	12.63	11.13	14.94	11.20	9.78	14.69	11.99	10.60
GUPNet	*ICCV21*	21.06	15.02	12.34	19.91	14.24	11.57	19.69	14.24	12.36	17.36	12.95	10.76	16.71	12.40	10.64	9.84	6.36	5.09
DID-M3D	*ECCV22*	22.75	15.52	12.61	22.19	15.96	12.86	22.42	15.30	12.43	21.40	14.79	12.05	20.56	14.07	11.88	21.92	14.79	12.10
DEVIANT	*ECCV22*	22.74	15.92	13.16	22.90	16.11	13.25	22.35	15.99	12.45	20.18	13.93	11.96	20.20	13.85	12.26	22.40	15.16	12.33
HomoLoss	*CVPR22*	14.31	12.27	11.12	19.32	13.26	11.51	18.23	13.19	12.56	17.69	13.01	12.23	16.33	13.40	10.76	15.88	13.89	11.42
CubeR-CNN	*CVPR23*	21.11	14.97	12.55	20.81	14.77	12.12	20.37	14.14	12.38	22.36	13.67	11.11	19.17	13.54	10.99	20.11	14.37	11.89
AEAM3D	-	23.13	16.03	13.19	23.24	16.28	13.35	23.08	16.01	12.98	23.06	15.77	12.92	21.31	15.40	12.52	22.55	15.70	12.80
Improvement		**+0.38**	**+0.11**	**+0.03**	**+0.34**	**+0.17**	**+0.10**	**+0.66**	**+0.02**	**+0.53**	**+1.66**	**+0.98**	**+0.87**	**+0.75**	**+1.33**	**+0.26**	**+0.15**	**+0.54**	**+0.49**



Low Light Heavy Rain Thick Fog

Fig. 3. Qualitative results on the validation set of hybrid dataset which contain all types of weather. We use green, red, blue boxes to denote ground-truth, our predictions and predictions of CubeR-CNN respectively.

Table 2. Comparison of our proposed method with the combinations of our base 3D detection network and popular enhancement models under various challenging conditions.

Scene	Methods	Venue	Car 3D@IOU=0.7		
			Easy	Mod.	Hard
Thick Fog	Trans	*CVPR22*	22.95	16.03	13.21
	MSBDN	*CVPR20*	20.11	14.14	11.55
	GCA	*WACV19*	21.21	14.08	12.49
	DCPDN	*CVPR18*	19.97	13.25	11.34
	Ours	-	**23.13**	**16.03**	**12.95**
Heavy Rain	Trans	*CVPR22*	20.29	13.89	11.67
	RESCAN	*ECCV18*	20.06	13.81	10.99
	VRGNet	*CVPR21*	21.55	12.98	11.01
	PRENet	*CVPR19*	20.11	13.34	10.67
	Ours	-	**23.06**	**15.77**	**12.92**
Low Light	Trans	*CVPR22*	14.7	10.53	9.18
	SCI	*CVPR22*	19.88	14.12	10.68
	IAT	*BMVC22*	19.84	13.59	10.94
	SID	*CVPR18*	17.78	12.21	10.32
	Ours	-	**22.55**	**15.70**	**12.82**

3.5. Qualitative Results

Figure 3 reveals the superior performance of AEAM3D compared to the current SOTA approach, CubeR-CNN[19], in three distinct environments. For example, in low light scenarios, where the environment is comparatively dark, GUPNet tends to miss objects, whereas AEAM3D accurately identifies nearly all objects. These observations highlight the considerable advantages of our method over other optimal approaches, that is, resilience against environmental challenges.

Table 3. Ablation study for the components of our method. Results are reported on hybrid datasets.

	Enc	Dec	3D@IoU=0.7		
			Easy↑	Mod.↑	Hard↑
(a)	✗	✗	18.53	13.09	10.89
(b)	✗	✓	19.12$_{\uparrow 0.59}$	14.43$_{\uparrow 1.34}$	11.74$_{\uparrow 0.85}$
(c)	✓	✗	20.35$_{\uparrow 1.82}$	14.86$_{\uparrow 1.77}$	12.11$_{\uparrow 1.22}$
(d)	✓	✓	23.22$_{\uparrow 4.69}$	15.55$_{\uparrow 2.46}$	12.31$_{\uparrow 1.42}$

4. CONCLUSION

In this study, we introduce AEAM3D, a monocular 3D object detection model incorporating an adaptive learning strategy and demonstrate exceptional performance in an array of challenging environments, including fog, rain, and low-light conditions. The adaptive learning strategy effectively regularizes the model, enabling AEAM3D to adapt to and perceive features across inclement weather conditions. This innovative approach substantially advances the practical applicability of monocular 3D object detection models. Extensive experimental results attest to the superiority of our proposed method over state-of-the-art approaches, both qualitatively and quantitatively, across various adverse environments.

5. REFERENCES

[1] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov, "Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15446–15456.

[2] Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Min-Jian Zhao, and Gim Hee Lee, "Rethinking iou-based optimization for single-stage 3d object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 544–561.

[3] Honghui Yang, Zili Liu, Xiaopei Wu, Wenxiao Wang, Wei Qian, Xiaofei He, and Deng Cai, "Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*. Springer, 2022, pp. 662–679.

[4] Xiaoke Shang, Gehui Li, Zhiying Jiang, Shaomin Zhang, Nai Ding, and Jinyuan Liu, "Holistic dynamic frequency transformer for image fusion and exposure correction," *Information Fusion*, vol. 102, pp. 102073, 2024.

[5] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang, "Delving into localization errors for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4721–4730.

[6] Xingyuan Li, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu, "Advmono3d: Advanced monocular 3d object detection with depth-aware robust adversarial training," 2023.

[7] Xingyuan Li, Jinyuan Liu, Yixin Lei, Long Ma, Xin Fan, and Risheng Liu, "Monotdp: Twin depth perception for monocular 3d object detection in adverse scenes," 2023.

[8] Garrick Brazil and Xiaoming Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[9] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2069–2078.

[10] 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*, 2019, pp. 8445–8453.

[11] Zhiying Jiang, Zengxi Zhang, Yiyao Yu, and Risheng Liu, "Bilevel modeling investigated generative adversarial framework for image restoration," *The Visual Computer*, vol. 39, no. 11, pp. 5563–5575, 2023.

[12] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang, "Single image dehazing via conditional generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8202–8211.

[13] Zhiying Jiang, Zhuoxiao Li, Shuzhou Yang, Xin Fan, and Risheng Liu, "Target oriented perceptual adversarial fusion network for underwater image enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6584–6598, 2022.

[14] Yanfu Zhang, Li Ding, and Gaurav Sharma, "Hazerd: an outdoor scene dataset and benchmark for single image dehazing," in *2017 IEEE international conference on image processing (ICIP)*, 2017, pp. 3205–3209.

[15] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1633–1642.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer.," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[17] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[18] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2147–2156.

[19] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari, "Omni3d: A large benchmark and model for 3d object detection in the wild," 2023.