FocalAngle3D: An Angle-Enhanced Two-Stage Model for 3D Detection

Bo Yang[1], Shengyin Jiang[1], Zeliang Ma[1], Dengyi Ji[2], Haiwen Li[1]

[1]Beijing University of Posts and Telecommunications [2]Beijing University of Technology

{bobyang,shengyin,mzl,lihaiwen}@bupt.edu.cn, 2973138766@qq.com

Abstract

Our model builds upon the StreamPETR paradigm. To improve the robustness of detection, particularly in challenging corruption scenarios, we first follow the 2D bounding box auxiliary supervision of RepDETR3D. Innovatively, our approach adds more precise depth information into the extracted 2d feature by integrating an intensive depth prediction network, which enhances feature extraction and contributes to more accurate 3D bounding box regression. Additionally, for orientation estimation, we encode the orientation angle with three phase-shifted encoding channels, proposed as Phase-Shifting Coder (PSC). Experimental results validate that incorporating more angle information via PSC enhances model robustness in capturing subtle angular variations and differences, consequently elevating orientation accuracy metrics.

1. Introduction

Accurate 3D detection of the surrounding environment is crucial in fields such as autonomous driving and robotics, which enables a more comprehensive understanding of the surrounding scene [1–6], including object translation, sizes, and velocity, and facilitates more effective action planning. In recent years, with advancements in computer vision and deep learning, significant progress has been made in the field of 3D detection. Currently, pure vision-based 3D detection can be categorized academically into four major classes: the LSS series, BEVFormer series [7], PETR series [8], and Sparse4D [9] series of 3D detection solutions.

The complexity of autonomous driving scenarios poses significant challenges to research in this field, and the robustness of existing 3D scene perception models in various challenging scenarios remains inadequately evaluated [10]. Under the pure visual detection paradigm, it fails to adequately address various complex lighting and weather conditions, such as low visibility scenarios in nighttime, rain, snow, or fog scenes.

To address robust detection in challenging scenarios, we adopt the efficient paradigm of 3D detection, StreamPETR [11], which inherently features strong spatiotemporal modeling and 3D detection performance. Building upon this foundation, we incorporate deformable attention to serve as spatial cross-attention and utilize 2D proposal loss for auxiliary supervision to optimize feature extraction networks following RepDETR3D. Furthermore, we add an intensive depth prediction network, integrating depth information supervision to enhance the positional accuracy of 3D detection boxes. To tackle image corruption in challenging scenarios, we propose a robust encoding method named PSC for orientation angle prediction, leveraging the diversity provided by additional angle information to better capture subtle angular variations and differences.

Our model achieved 3rd place in Track 1: robust BEV detection of the 2024 RoboDrive Challenge [12]. Simultaneously, it demonstrated significant performance on the validation set of nuScenes. This paper provides a detailed analysis of our model's performance on both datasets, highlighting its effectiveness in real-world scenarios.

2. Approach

2.1. One Stage Mono3D Head

2D Detection Network: Given that the current model encounters challenges in various complex lighting and weather conditions, such as low visibility scenarios in nighttime, backlit scenes, and adverse weather like rain, snow, or fog, where visual information in image features is prone to interference. To address this problem, we incorporate a one-stage 2D detection network YOLOX [13] following RepDETR3D. By employing a 2D detection task as auxiliary supervision, we enable the backbone network to capture better features.

YOLOX adopts the Anchor-Free paradigm, that is the prediction head predicts targets for each grid of the feature map and selects the targets with the highest peak confidence

Technical Report of the 2024 RoboDrive Challenge.
Track 1: Robust BEV Detection.

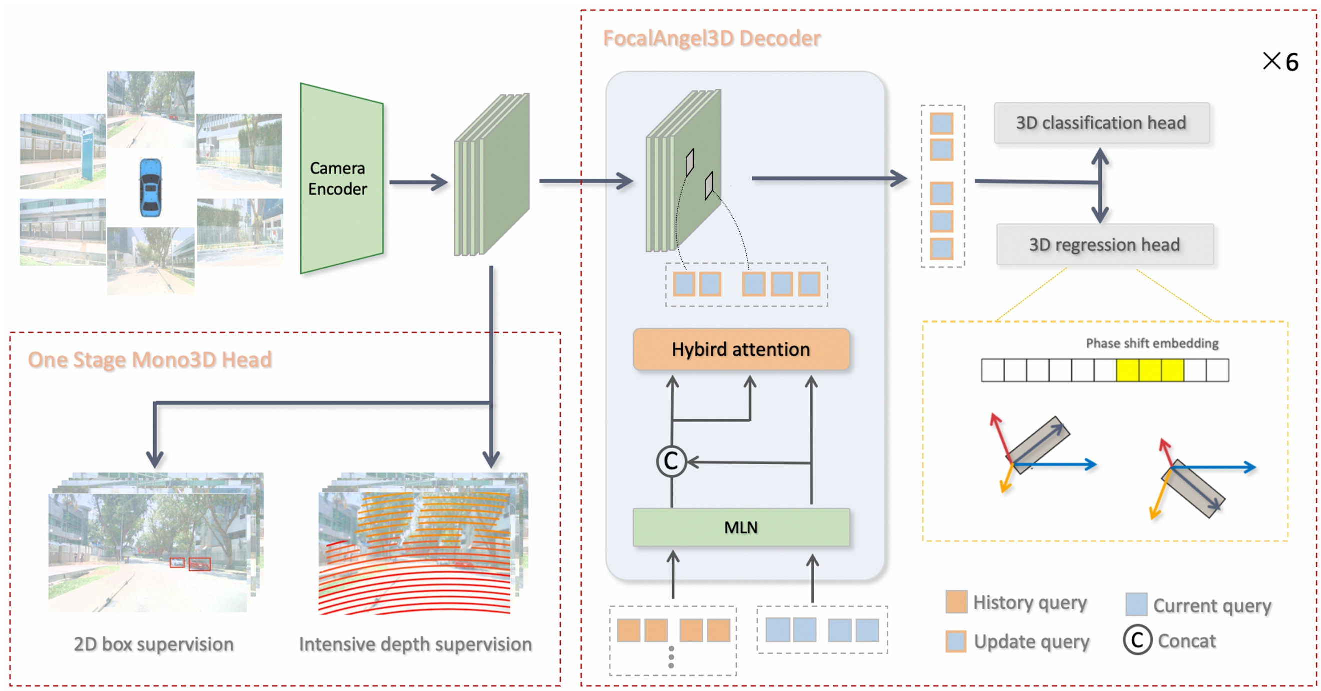

Figure 1. The main architecture of FocalAngel3D: surrounding images are input to the image encoder to extract images features, then firstly go through one stage mono3d prediction head, which consists of YOLOX[13] and our intensive depth network. The second stage is the FocalAngel3D decoder, the history queries, and the initialized queries are firstly input to the MLN for the temporal alignment, followed by a hybrid attention module. Deformable attention is implemented between the output of the updated current frame queries and image features, followed by the classification head and regression head respectively, where our phase shift coder method is used for angle prediction.

score as the final result. We benefit also from replacing the coupled detection head with a decoupled detection head, where Class, Regression, and IoU predictions are handled by three separate branches rather than a single 1x1 convolutional layer, which can enhance the network's convergence speed and accuracy [14]. After that, simOTA is used to assign positive predictions to GT. With the cost matrix and the dynamic number k of anchors required for each GT, the target anchors are selected.

Intensive Depth Head: The one-stage 2D detection network adopts the architecture of YOLOX [13]. To enhance the accuracy of 3D position prediction, we incorporated a depth prediction head into the existing architecture alongside the original class head, position head, and objectness head. This addition facilitates improved learning of depth features in the network. In existing mono3D paradigms for depth supervision, some depth prediction methods adopt the sparse paradigm, which supervises only the depth at the center of the 2D detection boxes. However, this sparse depth supervision method can only predict depth for a limited number of predicted box centers, resulting in insufficient convergence during early training stages. To mitigate this issue, we introduced dense depth prediction. Specifically, we conduct depth map prediction for the four feature maps with different shapes. We adapt 1×1 $Conv$ to output depth prediction values for these feature positions. Supervision is applied within the range of 60 meters, with gt depths spaced at intervals of 0.5 meters. We construct one-hot labels and the depth prediction loss is calculated using Binary Cross-Entropy (BCE) loss.

2.2. FocalAngle3D Decoder

Temporal Fusion: We align temporal queries to the current frame using the MLN structure following StreamPETR. Specifically, for historical queries, we initially align their 3D coordinates to the position of the current frame by ego-motion matrix. Subsequently, we apply an MLP layer to embed the ego-motion matrix, velocity, and time interval to r and β. Then, we utilize r and β to update the query and positional embedding. To ensure embedding consistency, we also employ the MLN to process the initialized queries for the current frame, with their velocity and time interval values set to 0. After that, we concatenate the memory query processed by MLN with the current query processed by MLN to a hybrid query. This hybrid query then acts as both keys and values and subsequently performs hybrid attention with the current query processed by MLN

Spatial Fusion: The output current query from the hybrid attention will be processed in the same method as PETR. Instead of using the original method of doing cross-attention between image tokens and queries, we use deformable attention, which projects the 3d object query onto the image features, helping to reduce the amount of computation and facilitating the interaction of the query with the relevant positional features.

2.3. Phase Shift Coder

In current BEV methods, the encoding method of angles commonly utilizes sine-cosine encoding [7, 8]. However, this encoding method only utilizes two values, which are $sin(x)$ and $cos(x)$, thereby being limited by dimensionality in representing angles. While these two values can provide sufficient information to represent angles, they may restrict the model's ability to capture subtle variations in angles in certain scenarios.

Therefore, in our competition, we introduced a novel angle encoding method called phase-shift encoding. This method utilizes three values,which are $cos(x)$, $cos(x + 120)$ and $cos(x + 240)$. In contrast to the two values used in sine-cosine encoding, phase-shift encoding provides more angle information. This diversity enables the model to better capture subtle variations and differences in angles. The three values are phased at intervals of 120 degrees, providing additional constraints and assistance, making it easier for the model to distinguish differences between different angles.

During training, for each bounding box, our angle prediction head predicts the values of three encodings. For each ground truth angle phi, which ranges from $-\pi$ to π, we encode the ground truth using the following equation and calculate the L1-loss with the predicted values.

$$\begin{aligned} x_1 &= \cos(\varphi + 120^\circ) \\ x_2 &= \cos(\varphi + 240^\circ) \\ x_3 &= \cos(\varphi) \end{aligned} \tag{1}$$

During inference, the output of the angle prediction head is the three phase-shift encoding values predicted by the network. At this stage, the predicted values can be decoded using the following equation to obtain the actual rotation angle, where $Nstep = 3$ represents 3 steps of phase shift.

$$\varphi = -\arctan \frac{\sum_{n=1}^{N_{\text{step}}} x_n \sin\left(\frac{2n\pi}{N_{\text{step}}}\right)}{\sum_{n=1}^{N_{\text{step}}} x_n \cos\left(\frac{2n\pi}{N_{\text{step}}}\right)} \tag{2}$$

In theory, this phase-shift encoding can be adjusted and modified Flexibly. By altering the phase difference, the encoding dimension can be changed to adapt to different problems and scenarios. For example, $Nstep$ can be adjusted to 5 steps, 7 steps, or other odd numbers. This flexibility enhances the versatility and adaptability of phase-shift encoding. Since PSC encoding is derivable, we can insert the decoding process directly into the forward structure of the network during training without using encoding. If this approach is adopted, PSC becomes part of the network's forward structure.

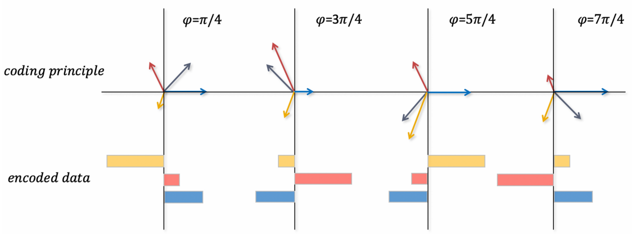

Figure 2. Phase shifting coder: we show the change of the coded principal of four different angles and the corresponding coded values. Yellow, red, and blue bars correspond to x1, x2, and x3.

3. Experiments

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [12] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [1] (which contains 700 scenes) and tested on the held-out competition evaluation sets (which contains 150 scenes). The evaluation data was created following RoboDepth [15–17], RoboBEV [10, 18, 19], and Robo3D [20, 21]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. For more details, please refer to the corresponding GitHub repositories.

3.2. Experiment Settings

The focalAngle3D network is implemented using the PyTorch framework [22] and is based on the MMDetection3D codebase [23]. 8 NVIDIA A100 GPUs are used in our model training experiments, each with a batch size of 2. We use EVA02-L [24] as the backbone of the model, which is pretrained on the COCO dataset [25]. Adam is used as the optimizer, the learning rate of the model is set to $4e-4$ and the learning rate of the backbone network is set to 0.1 times the base learning rate. CosineAnnealing is used as the learning rate strategy, and the warmup phase is the initial 500 iterations.

3.3. Ablation Study

As listed in Table 1, ablation experiments show that after using the intensive depth network supervision, the image feature with better depth information is learned in the mono3d stage, which helps to regress more accurate 3d box positions and more accurate speed in FocalAngle3D decoder, as reflected in the 0.01 improvement in mATE, and 0.01 improvement in mAVE. After using the PSC, the mAOE metric

Table 1. FocalAngle3D results on nuScenes val dataset.

Method	*depth*	*psc*	*NDS*↑	*mAP*↑	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓
$FocalAngle3D$			0.613	0.531	0.529	0.255	0.296	0.240	0.201
$FocalAngle3D$	✓		0.618	0.537	**0.518**	0.254	0.299	**0.228**	0.206
$FocalAngle3D$	✓	✓	**0.620**	0.536	0.529	0.251	**0.276**	0.224	0.196

Table 2. FocalAngle3D results on corruption test dataset.

Method	*NDS*↑	*mAP*↑	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓
$FocalAngle3D$	0.490	0.436	0.613	0.375	0.441	0.521	0.328

improves by 0.02 compared to the baseline model, reflecting the improvement in angle prediction by this coding way, and this enrichment of coded information allows the model to better capture subtle changes and differences in angle.

4. Conclusion

In order to improve the robustness of the model, we improved the network in the competition by adding an intensive depth prediction network to help the network extract the 2D features with more accurate depth positions, which is beneficial to the accuracy of the 3D box position regression. For the angle prediction of the bounding box, we propose to use the angle coding paradigm of the PSC. Experimental results also prove that our improvement has significant improvement on the model to predict the position, angle, and speed of the 3d bounding boxes.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[2] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[3] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[4] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[5] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[6] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[7] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[8] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[9] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[10] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

[11] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[12] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding,

Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The robodrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[14] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023.

[15] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Benoit Cottereau, Lai Xing Ng, and Wei Tsang Ooi. The robodepth benchmark for robust out-of-distribution depth estimation under corruptions. https://github.com/ldkong1205/RoboDepth, 2023.

[17] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Ding Zhao, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The robodepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[18] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[19] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robobev benchmark for robust bird's eye view detection under common corruption and domain shift. https://github.com/Daniel-xsy/RoboBEV, 2023.

[20] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[21] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robo3d benchmark for robust and reliable 3d perception. https://github.com/ldkong1205/Robo3D, 2023.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[23] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[24] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.