An Effective Motorcycle Helmet Object Detection Framework for Intelligent Traffic Safety

Shun Cui[1], Tiantian Zhang[1,2†], Hao Sun[*], Xuyang Zhou[1],
Wenqing Yu[1], Aigong Zhen[1], Qihang Wu[3], Zhongjiang He[1]
[1]China Telecom Corporation Ltd. Data&AI Technology Company
[2]Beijing University of Posts and Telecommunications
[3]National University of Singapore
{cuis2, zhouxy26, yuwq, zhenag, hezj}@chinatelecom.cn, e1010952@u.nus.edu, iszhang_tt@bupt.edu.cn

Abstract

Detecting violations of motorcycle helmet rules is an important computer vision task that can greatly protect the lives of motorcycle drivers and passengers in traffic accidents. This abnormal event detection problem can be viewed as an image object detection task, which aims to detect the location of the motorcycle driver and passenger in the image and whether they are wearing helmets. In this paper, we propose a motorcycle helmet object detection (MHOD) framework to achieve this task. Specifically, we first utilize the object detection network with ensemble model to predict the location and category of all objects in videos which can improve the accuracy and robustness of detection model. Then for the scarcity of passenger category training data, the Passenger Recall Module (PRM) is designed via tracking refinement which greatly improves passenger category recall. Finally, we introduce the category refine module (CRM) to correct the category by combining the temporal information in the video. On the test dataset of AI City Challenge 2023 Track5, we achieve significant result compared with other teams, the proposed model ranks first on the public leaderboard of the challenge.

1. Introduction

Motorcycles remain a popular means of transportation, especially in developing countries like India. Compared with conventional vehicles, motorcycle riders are at a higher risk of accidents due to the limited protection offered by their vehicles. As a result, motorcyclists are required by law to wear helmets. The automatic detection of motorcyclists without helmets is a crucial task for enforcing strict regulatory traffic safety measures. This computer vision task holds immense potential to safeguard the lives of motorcycle drivers and passengers in traffic accidents.

The detection of motorcycle riders and passengers wearing helmets is a challenging example of object detection in computer vision research. This task involves identifying and localizing the motorcycle vehicle, its riders, and passengers, and determining whether they are wearing helmets. Object detection is a crucial computer vision task that involves recognizing and localizing objects of interest in images and video.

In recent years, there has been a development in object detection from CNN-based methods to transformer-based methods. The main reason for this development is that CNN-based methods rely on local patterns in the image and are limited by the size and complexity of the receptive field. On the other hand, transformer-based methods capture global contextual information in the image by exploiting self-attention mechanisms, which helps to identify objects more accurately and efficiently. Recent DETA [13] shows a simpler alternative training mechanism of transformer-based compared to other detection transformer method. This paper uses DETA as detector to detect motorcycle and passengers.

However, the study of such a task poses several challenges. For example, it is difficult to accurately detect helmet use under different lighting conditions and camera angles. In addition, in traffic monitoring systems, cameras are often located at high altitudes, resulting in low resolution video. As Fig. 1 shows, lighting, weather, blur, etc. challenges the given training set of 2023 AI City Track 5. In order to cope with these complex scenarios and to improve the robustness of the model, we will use the model ensemble strategy, the details will be described in Sec. 3.2. Ensemble modeling has been shown in previous studies [4] to achieve

[*]Corresponding authors, China Telecom Corporation Ltd. Data&AI Technology Company, sunh10@chinatelecom.cn
[†]This work is done when Tiantian is an intern at China Telecom Corporation Ltd. Data&AI Technology Company

Figure 1. The scenes in the visualization part of the video, from left to right are night, fog, and background chaos, in that order.

more robust object detection and higher accuracy. Using this approach, we aim to achieve more accurate results.

As shown in the Tab. 1, we count the number of boxes in each category in the given training dataset, the category imbalance problem is very serious, especially the lack of Passenger 2 data. Analyzing the dataset, we see that Passenger 2 only appears in two videos, 005 and 091. We visualized the sample in Fig. 2, it is observed that the Passenger 2 in 005 is a small child in front of the motorcycle, and the model struggles to classify this instance. On the other hand, the person identified as Passenger 2 in 091 is at the rear of the motorcycle, and this situation is more consistent with our perception. To solve the lack of Passenger 2 sample, we develop a Passenger Recall Module (PRM) based track which introduced in Sec. 3.3 to further refine detection bounding box obtained by detector. Moreover, inspired by the idea of tracking, the corresponding box of a track ID should not switch categories during movement. We use the SORT [2], which is a multiple object tracking (MOT) algorithm, to correct the category information of the corresponding prediction frame of the same tracking target appearance frame. The details are described in Sec. 3.4.

In summary, the main contributions of this paper are summarized as follows:

- We propose a robust Motorcycle Helmet Object Detection (MHOD) framework with model ensemble.
- We present the Passenger Recall Module (PRM), which significantly improves passenger category recall.
- We introduce the Category Refine Module (CRM) to improve accuracy, leading to improvements in motorcycle and passenger detection.
- We show that the proposed framework achieves first place in the AI City 23 Challenge Track 5 final leaderboard results with a score of 0.8340.

Class Id	Class Name	Instances
1	motorbike	31121
2	DHelmet	23220
3	DNoHelmet	6856
4	P1Helmet	94
5	P1NoHelmet	4280
6	P2Helmet	0
7	P2NoHelmet	40

Table 1. Category statistics for all targets in the training set.

Figure 2. Image visualization of Passenger 2.The left and right of the image respectively Passenger 2 appeared in the video 005 and 091.

2. Related Work

2.1. Object Detection

As one of the fundamental computer vision problems, object detection aims at locating and classifying objects of interest in images or videos, and labeling them with rect-

angular bounding boxes. There is an abundance of literature on object detection methods in computer vision, including conventional iterative models (*e.g.* RCNN [7], Fast R-CNN [6], and Faster R-CNN [15]) as well as modern deep learning-based frameworks (e.g. YOLO [14] and SSD [10]). In particular, we use Detectron2 [17] for PRM in this paper because it is lightweight and efficient. Detectron2 is a Facebook AI research library that provides advanced object detection and segmentation algorithms.

In addition to the above methods based on convolutional networks, the transformer architecture has been recently applied for object detection as backbone, which achieves an impressive speed-accuracy trade-off, such as Vision Transformer (ViT) [5] and Swin Transformer [11]. The first transformer-based detector proposed in 2020, Detecting Objects with Transformers (DETR) [3], surpassed state-of-the-art accuracy on the COCO dataset. Since then, several variants and alternatives have been proposed, such as Deformable DETR [19], a transformer-based detector that uses deformable attention to address issues related to small object detection and imbalanced data. These models directly transforms queries to unique objects by using one-to-one bipartite matching during training. DETA [13] proposes training with IoU-based label assignment rather than vanilla one-to-one matching of detection transformers. The assignment strategy can effectively improve the mAP.

2.2. Multiple Object Tracking

Multiple object tracking (MOT) is the important area in computer vision. The task of MOT is to detect multiple targets such as pedestrians, cars, animals, etc. in the video and assign IDs for trajectory tracking, without knowing the number of targets in advance. With the recent progress in object detection, tracking-by-detection [1] has become the de facto approach to multiple object tracking. It consists of first detecting the objects in the individual frames and then associating these detections with trajectories, known as tracklets. Simple Online and Realtime Tracking (SORT) [2] is a pragmatic method to MOT with a focus on simple, effective algorithms. It performs Kalman filtering in image space and frame-by-frame data association using the Hungarian method with an association metric that measures bounding box overlap. In this paper, we fuse the SORT results to determine the object class, which can improve the accuracy of some difficult object class during trajectory tracking.

3. Method

3.1. Overview

An overview of our Motorcycle Helmet Object Detection (MHOD) framework is presented in Fig. 3. Generally, it consists of three parts. First, our proposed framework adopts ensemble technique to improve the performance. In the second part, the Passenger Recall Module (PRM) is performed to improve passenger category recall. The third part is the Category Refine Module (CRM), which aims to reduce the number of class switches in the same trajectory. All the modules and components are described in detail in the following sections.

3.2. Ensemble Model

Since the complex variability and low resolution of video scenes, our proposed framework model ensemble with different initialization processes to improve performance. The object detection approach used in this paper is based on the recent DETA algorithm [13]. DETA shows a simpler alternative training mechanism of detection transformer compared to recent detection transformer methods [3, 18]. This alternative enjoys a significant advantage in training efficiency, especially with a short training schedule. We fetch the bounding box of the detected object in each video frame and the corresponding confidence using the detection model:

$$B_i = \{(b_i, f_i)|i \in v\}, \tag{1}$$

where b_i is the corresponding bounding box information, f_i is the time frame, and v is the frame length of video. After getting the detection results, we have a bounding box $b = (cls, x_c, y_c, w, h, s)$, where cls is the class id of bounding box,(x_c, y_c) is the position of center point, (w, h) is the width and height of bounding box, and s is the confidence score. We perform non-max-suppression(nms) to filter detection boxes that overlap the same objects. Accordingly, the final prediction extracted from the two individual models by using nms is generally formulated as follows:

$$Z_i = \{nms(B_{E1,i}, B_{E2,i})|i \in v\}, \tag{2}$$

where Z represents the final prediction. Both E_1 and E_2 are DETA models fine-tuned on AI City Challenge dataset.

3.3. Passenger Recall Module

Based on the statistical results presented in Tab. 1 for the training set, the sample size of Passenger 2 proves to be unusually small and atypical. Thus, post-processing techniques are applied to refine the detection bounding boxes for Passenger 2.

We employ the open source framework Detectron2 [17] pretrained on the COCO dataset [9] to obtain the collection of person bounding boxes $P = \{p_1, p_2, p_3, \cdots\}$, where $p = \{x_c, y_c, w, h, s, f\}$. We obtain the collection of motorcycle bounding boxes $M = \{m_1, m_2, m_3, \cdots\}$ from Z. For each $m_i \in M$, when it satisfies the following condition, record all $p_j \in P$ that match with m_i:

$$\sum_{p_j \in P} \mathbb{I}[iou(m_i, p_j) > \alpha] \geq 3, \tag{3}$$

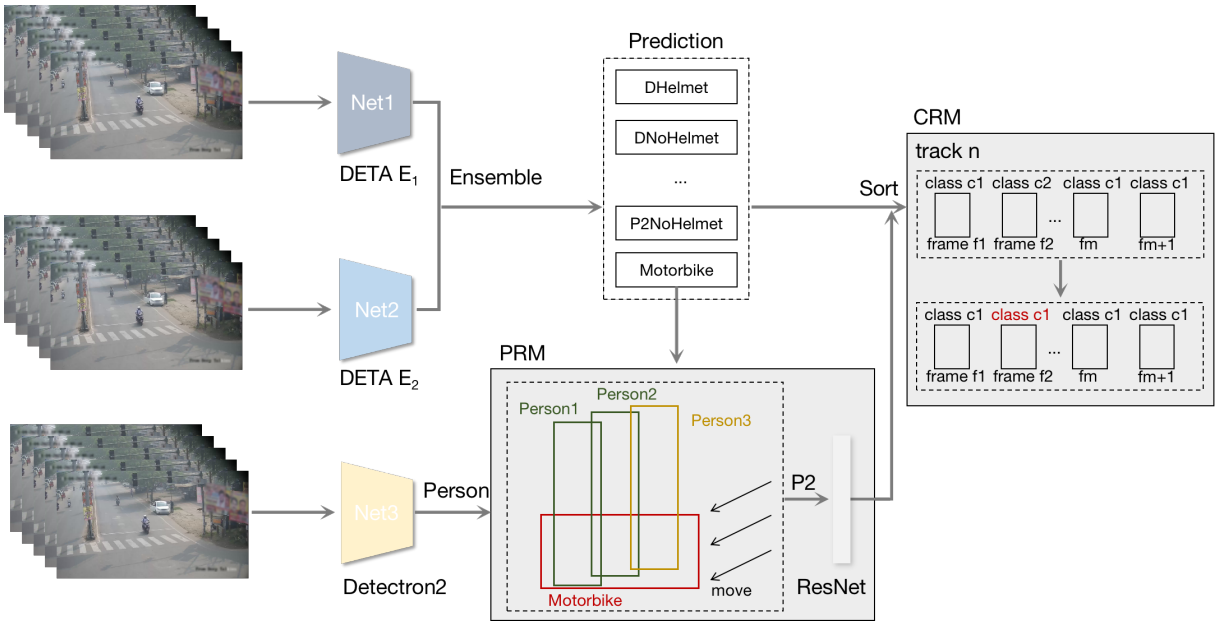

Figure 3. An overview of MHOD framework architecture. We first use DETA with model ensemble strategy for detection, then the detected motorcycle boxes and person bounding boxes detected by the detetron2 model are sent to PRM for Passenger 2 boxes recall. Subsequently, the recalled Passenger 2 boxes are sent to resnet for classification to determine whether the person is wearing a helmet or not. Finally, the recalled Passenger 2 boxes and the output of the detection model are sent to CRM for category correction.

where α is the coefficient that controls the size of the iou, $iou(x, y)$ represents the intersection over union (IoU) between bounding boxes x and y. $\mathbb{I}(\cdot)$ denotes the distinguish function, returning 1 when under some conditions.

We use SORT [2] to predict the track of the person box and record the motion direction of the person. We compute the motion direction of each bounding box based on the correlation between consecutive frames, which can be expressed as $arctan(\frac{dy}{dx})$, where dx and dy denote the displacement of the center coordinates of the box from the previous frame to the current frame.

Then we determine the position of Passenger 2 according to the track direction, i.e., Passenger 2 is the last person box in the track direction. We further train a binary classification network on the training set to determine whether a pedestrian is wearing a helmet. We pass the box corresponding to Passenger 2 obtained previously to the classification network to determine whether a helmet is present.

3.4. Category Refine Module

In the video, we found that as the motor vehicle drives out of the camera, the label predicted by the model will change as the target gradually becomes smaller. Inspired by the idea of tracking, the corresponding box of a track ID should not switch categories during movement. SORT [2] is one of the most typical methods of tracking-by-detection category. We apply SORT to associate the detected objects of different frames in test videos, and get the trajectories of motorcycle and person. The trackers use information matrices of the bounding box to assign corresponding tracklet IDs with person and motorcycle detection. Finally, the tracker generates a set tracklets:

$$T_{id} = \{(id, b_i, f_i) | i \in v\}, \tag{4}$$

where T_{id} is the tracklet corresponding to ID. For a given tracking ID T_{id}, we compute the frequency of detected classes across all frames. When the frequency of a certain class c is greater than 50% of the total number of detections for the given tracking ID, we consider it to be of class c and refine the class label across all frames accordingly.

4. Experiments

4.1. Datasets

The Track 5 dataset in 2023 AI City Challenge [12] consists of a training set and a test set, each containing 100 videos. Each video is recorded at a resolution of 1920×1080, with a duration of 20 seconds and a frame rate of 10 fps. In the training set, the groundtruth bounding boxes of motorcycle and motorcycle rider(s) with or without helmets are provided. Each motorcycle in the annotated frame has bounding box annotation of each rider with or without helmet information, for upto a maximum of 3 riders (i.e., driver, passenger 1, Passenger 2) in a motorcycle. The purpose of the challenge is to devise an algorithm that

Approach	Score
Baseline (DETA)	0.5259
Baseline + Ensemble	0.6973
Baseline + Ensemble + PRM	0.8333
Baseline + Ensemble + PRM + CRM	0.8340

Table 2. Comparison of the influence of different modules on the detect performance. The proposed PRM leads to the most significant improvement. The use of the combined PRM and CRM which is based track completion component brings lower improvements.

is capable of identifying motorcycle and motorcycle rider(s) with or without helmet. Similar to the training dataset, each rider in a motorcycle is to be separately identified if they have a helmet or not.

4.2. Implementation Details

Our solution is implemented with the Pytorch framework. We choose DETA [13] with Swin-L backbone [11] as the detection network, Resnet18 [8] as the classification network, and SORT [2] as the tracking algorithm.

Training phase. Experiments are executed with batch size 4 on 4 Nvidia V100 GPUs. The model is fine-tuned on AI City Challenge dataset with Adam optimizer for 8 epochs. Learning rate is 5e-6 and weight decay is 1e-4. Intermediate size of the feedforward layers in the transformer blocks is 2048 and number of attention heads inside the transformer's attentions is 8. The image scale jittering during training is randomly selected from [720, 768, 816, 864, 912, 960, 1008, 1056, 1104, 1152, 1200] pixel for the shorter side, and the long side does not exceed 2000 pixel. The short side is 1200 pixel at the time of testing. Moreover, we load the pre-trained model parameters on Objects365 [16]. The two models we used for integration differ only in terms of the queries used during initialization, which are set to 300 and 900, respectively. We adopt ResNet-18 [8] pre-trained on ImageNet as classification model, and the input is the person bounding boxes with a resize of 256×192. We use the person bounding boxes in the AI City Challenge for training, and the ratios for training and testing are 0.9 and 0.1, respectively. The learning rate is 0.04 and the weight decay is 5e-4 for training 100 epochs with the learning rate decay strategy of CosineAnealingLR.

Test phase. Firstly, we extract frames from video clips in test set with python script, then use the model fine-tuned on AI City Challenge dataset to detect the motorcycle and person. The coefficient α is set to 0.175. For Detectron2 [17], we load the mask_rcnn_X_101_32x8d_FPN_3x.yaml setting to detect person. The settings related to the tracking algorithm SORT [2] are as follows, the minimum number of associated detections before track is initialised to 0, and the maximum number of frames to keep alive a track without associated detections is set to 10.

Figure 4. Example visualization of passenger recall.

4.3. Metrics of Evaluation

The challenge ranking is based on the mean Average Precision (mAP) score across all frames in the test videos. mAP measures the mean of average precision (the area under the Precision-Recall curve) over all the object classes.

4.4. Experiments Results

Ablation Study. In Tab. 2, we investigate how several components contribute to the final results. We observe that adopting DETA with ensemble model method outperforms baseline DETA model by 17.14%. It also appears that the PRM can noticeably improve the performance. In Fig. 4, the direction of motion of this sample is down to the right, and this case marks the third person as Passenger 2 to be fed into the classification model to determine whether or not to wear a helmet.

Furthermore, our track refinement approach indeed assists with CRM. The results show that post-processing resulting tracks using prior knowledge further enhances the results. We perform tracking visualization on the training set 090 video, and it can be seen that the object prediction category with id 42 at frame 12 of Fig. 5a is DHelmet, but at frame 24 of Fig. 5b, it is predicted to be P1NoHelmet, which is modified to DHelmet using the proposed CRM strategy.

Comparison with other teams. The proposed system is submitted to the Track5 of AI City Challenge 2023 for evaluation. As shown in Tab. 3, our method surpasses these methods by a large margin with score 0.8340 and ranks first place among over 30+ teams from all over the world.

5. Conclusion

In this paper, we propose the Motorcycle Helmet Object Detection (MHOD) framework to detect motorcycle helmets. The MHOD module employs an object detection network to predict the location and class of all targets in the video. We introduce the passenger recall module (PRM)

(a) fram12 (b) fram24

Figure 5. Visualization of category refinement results. The red box text "-" with id 42 in the (b) represents the category corrected using the PRM strategy in the front, and the predicted category in the back.

Rank	Team ID	Team Name	Score
1	**58**	**CTC-AI (ours)**	**0.8340**
2	33	SKKU Automation Lab	0.7754
3	37	SmartVision	0.6997
4	18	UT_He	0.6422
5	16	UT_NYCU_SUNY-Albany	0.6389
6	45	UT1	0.6112
7	192	Legends	0.5861
8	55	NYCU - Road Beast	0.5569
9	145	WITAI-513	0.5474
10	11	AIMIZ	0.5377

Table 3. Top 10 Leaderboard of Track5 in the AI City Challenge 2023.

for tracking refinement to improve passenger category recall, and the category refinement module (CRM) to correct the object categories. The PRM is an extensible module that mainly implements recall for Passenger 2 in this paper. We can improve the effectiveness of this framework by developing strategies appropriate for Passenger 1 in future work. Experiments results on the public test set of 2023 AI City Challenge Track5 demonstrate the effectiveness of our method, which achieves score of 0.8340, ranking first on the leaderboard.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2, 3, 4, 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[4] A. Casado-García and J. Heras. Ensemble methods for object detection, 2019. https://github.com/ancasag/ensembleObjectDetection. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mansur Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9978–9988, 2020. 3

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE*, 2016. 5

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2021. 3, 5

[12] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 4

[13] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back, 2022. 1, 3, 5

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[16] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 5

[17] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 5

[18] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 3

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. 3