C4MTS: Challenge on Categorizing Missing Traffic Signs from Contextual Cues

Varun Gupta[1†], Sandeep Nagar[1], Suman Paul Choudhury[3], Rohit Singh[3],
Anandita Jamwal[2], Vibhu Gupta[2],
Anbumani Subramanian[1†], C.V. Jawahar[1†], and Rohit Saluja[2†]
https://cvit.iiit.ac.in/ncvpripg2023/c4mtschallenge

[1] IIIT Hyderabad
[2] IIT Mandi
[3] SAHAJ AI, India

Abstract. Traffic signs, despite being crucial for road safety, frequently remain absent. This challenge provides 200 scenes from a recent *Missing Traffic Signs Video Dataset* (MTSVD), distributed over four types of missing traffic signs: *left-hand-curve*, *right-hand-curve*, *gap-in-median*, and *side-road-left*, individually observed with their respective contextual cues. 2000 training images, each containing one of the four traffic signs with corresponding bounding boxes, are provided. Two tasks are proposed for the challenge: i) *Object Detection*, wherein the model is trained using bounding box annotations, and ii) *Missing Traffic Sign Scene Categorization*, wherein the model is trained using road scene images with in-painted traffic signs, provided with the challenge dataset. Baselines were provided to the participants for both tasks.54 teams registered for the challenge. Overall, the participants could improve the top-1 accuracy significantly by a margin of 31.5% over the baseline. This work presents the MTSVD in detail, challenge baselines, and the methodology undertaken by the top 2 teams.

Keywords: Missing Traffic Signs, Missing Objects, Context, Object Detection, Scene Categorization.

1 Introduction

Global Autonomous Vehicle (AV) market is projected to reach 2162 Billion by 2030 [1]. Around 1.3M lives are lost yearly in road accidents [2], with traffic sign violations contributing significantly to this number [3]. The traffic signs are generally installed at the side of the road to control traffic flow or convey information about the road environment to Vulnerable Road Users (VRUs). Often, the information is also available in the form of cues present in the context around the traffic signs (*obstacle-delineator* in Fig. 1) or in the cues away from it (*pedestrian-crossing* in Fig. 1), which we refer to as *contextual cues*.

[†]Challenge organizers.
Joint report with coordinators and invited contributions from the challenge winners.

Fig. 1: MTSVD [4] samples having traffic signs with contextual cues. **Anti-clockwise from top-left**: a *left-hand-curve*, *obstacle-delineator*, *gap-in-median*, and *right-hand-curve*. The red markings are added for visual emphasis.

The Missing Traffic Signs Video Dataset (MTSVD) [4] is the first publicly accessible data for missing traffic signs. The dataset contains 4000 missing sign intervals, the first of its kind, covering more than 500K frames across 30 traffic sign categories for which visual context exists. The MTSVD further contains bounding-box annotations of existing traffic signs on the road along with their context-interval markings, if present. Exact details are quantified and visualized in Section 2.

The C4MTS challenge of Categorizing Missing Traffic Signs provides 200 natural scenes from MTSVD uniformly distributed over 4 types of missing traffic signs. The traffic signs are common but individually observed with contextual cues (see samples in Fig. 1). In the given examples, contextual cues like *rumble strips*, *side roads*, etc., are present, but the traffic signs are missing, perhaps due to improper planning, etc. To learn the relationship between the traffic signs and the contextual cues, 2000 training images (similar to those in Fig. 1 with single traffic signs), each containing one of the 4 traffic signs (commonly and individually visible with contextual cues) and corresponding bounding boxes are also provided. The following tasks are proposed for the challenge: i) Object Detection and ii) Missing Traffic Sign Scene Classification. The details of the individual tasks are given in Section 4.

2 Missing Traffic Signs Video Dataset

Prior missing object datasets either had the cue in their immediate vicinity or had a consistent relationship between the missing object and the corresponding

cue [5] [6]. However, the MTSVD contains multiple cue contexts and complementary cue-object relationships, making MTSVD the most challenging and diverse, publicly accessible missing objects dataset. To ensure privacy, human-faces and vehicle-license-plates have been blurred in the dataset using [7].This section primarily describes the dataset in detail.

Fig. 2: Frame-level traffic sign annotations in the MTSVD [4]. The dataset contains bounding-box track annotations with multiple attributes like *tilted*, *occluded*, etc., and is captured in diverse weather and lighting conditions.

2.1 MTSVD in Numbers

The various traffic signs present in the dataset are illustrated as a word cloud in Fig. 3. The traffic signs in the MTSVD are spread across 74 types of traffic signs like *signboard*, *speed-limit*, and *chevron-left/right*, covering the significant spread. Sample annotations from the MTSVD are exhibited in Fig. 2. Fig. 5a illustrates the top-30 most frequent traffic signs in the dataset by the number of tracks annotated. *Signboard* primarily contain, but are not limited to, the overhead direction markers on the roads and other informatory markings, for example, *CCTV Surveillance*, *National-Highway* marking, etc. Further, *other* category comprises the signs that either do not fit well in the rest of the types or are challenging to define precisely. Moreover, MTSVD contains context frame annotations for various traffic sign categories, which can be identified by the corresponding scene context and are made available for existing traffic signs and, importantly, missing traffic signs. Fig. 5b highlights the missing signs categories and the corresponding interval counts identified. Finally, Table 1 mentions a precise qualitative overview. pie chart containing information about the categories in the MTSVD split into two halves: context and no-context.

Fig. 3: Various context categories in the MTSVD, where the corresponding traffic sign remains absent in the scene. The contexts may slightly differ from one scene to another, as it's a real-world artifact, but they remain fairly consistent and follow an observable pattern.

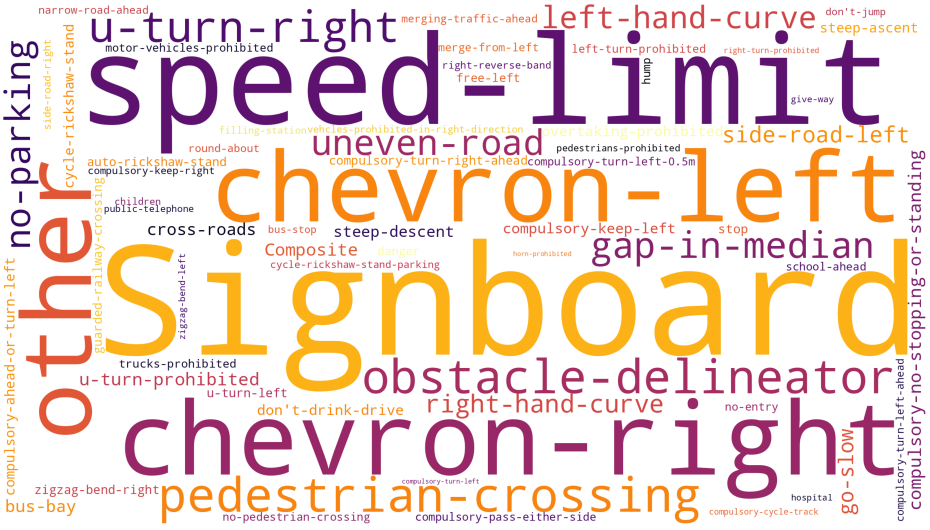
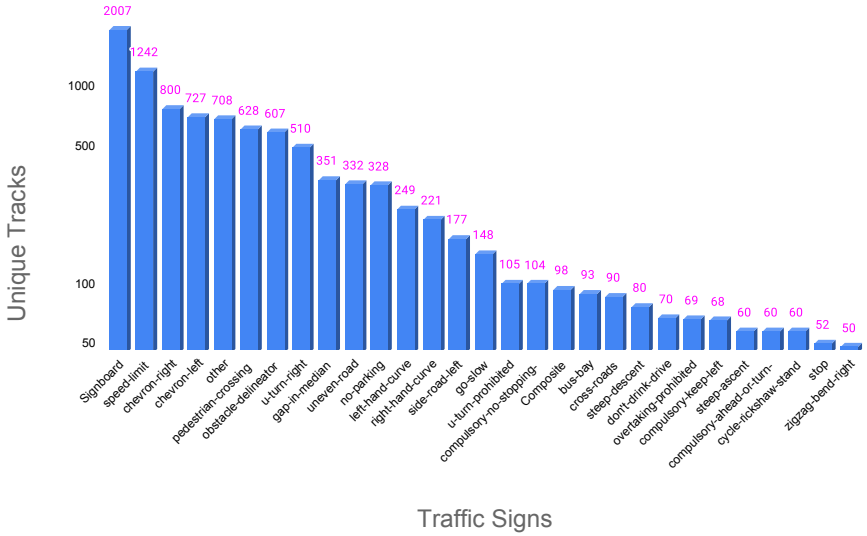
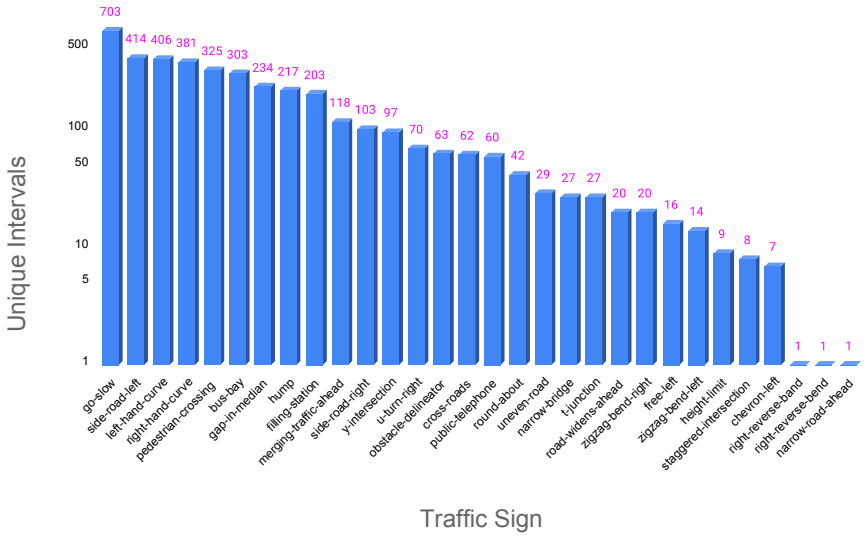


Fig. 4: Word-cloud of the traffic sign categories in the MTSVD [4].



(a) Top 30 of the most frequent traffic signs in the MTSVD by the number of unique tracks annotated.



(b) Missing traffic sign categories marked in the MTSVD. The annotations are made available as frame intervals, where the context for the corresponding *missing* traffic sign remains visible

Fig. 5: Annotated and Missing traffic signs in the MTSVD [4]

Table 1: Overview of the MTSVD. † Each frame has dense attributes for shape, category, tilt, etc., along with the bounding box and belongs to a corresponding track, identified by a unique video-level track ID.

Type	**Count**
Unique Videos	1,590
Traffic Sign Categories	74
Missing Signs Categories	30
Unique Traffic-Sign Tracks	10,510
Annotated Frames †	1,200,0658
Existing Traffic Signs Context Intervals	3,470
Missing Traffic Sign Context Intervals	3,977
Missing Traffic Sign Context Frames	573,441
Existing Traffic Sign Context Frames	839,172

3 Challenge Impact and Significance

The challenge is proposed to boost the research on missing objects, which is exciting but is being studied by only a handful of groups. Till now, the problems related to the missing object are studied for identifying missing barricades [5], missing curbs [6], or locating potential positions for non-existing pedestrians [8]. Such works and related models are specific to the task and do not involve multiple classes. However, missing traffic signs is a reality and of great practical importance in the Indian scenario and similar countries lacking road infrastructure. Moreover, the variety of traffic signs that are found missing is more than the previous missing object-related works. Classifying scenes with missing traffic signs can help the government install them at appropriate locations, enable vehicle companies to include dash-cams with adequate warnings even in case of missing signs, and have the potential to help autonomous vehicles follow applicable rules even in the absence of traffic signs. Thus, it can significantly impact the industrial community, government, and smart-city planners.

4 Participation Rules and Evaluation Criteria

The challenge consists of the two tasks mentioned below:

- **Task 1**: Traffic Sign Detection: Localization and classification of traffic signs in the road scenes.
- **Task 2**: Classifying missing traffic sign scenes.

Both tasks involve four types of traffic signs: *left-hand-curve, right-hand-curve, gap-in-median*, and *side-road-left*. These traffic sign types are chosen based on two factors; firstly, their frequency is high among the missing traffic sign scenes in the MTSVD dataset [4]. Secondly, they are commonly observed without interference from other traffic signs or contextual cues, possibly avoiding overcomplicating the task and encouraging more participation.

4.1 Evaluation Criteria

Separate data and evaluations are proposed for the two tasks. For the first task, 2000 in training images and 200 in validation images will be provided. 200 test images will be provided before the end of the competition. The mean Average Precision (mAP) score will be considered for evaluating the different detection methods. To train the models for task 2, the 1000 images with in-painted traffic signs will be provided, in-painted using a SoTA in-painting method [9]. To evaluate the methods for task 2, a validation set and a test set of 200 images each are provided (test set before the end of the competition). In both sets, 100 frames are from actual missing traffic sign video sequences, and 100 frames are from images with traffic signs in-painted are used. The actual missing sign images are included to encourage participants to create in-painting agnostic solutions. Standard error rates will be considered to evaluate different methods for the classification task [10].

4.2 Participation Rules

The participation rules are established as follows:

1. Any individual or group can use email to create a participation ID for the competition.
2. Making multiple IDs of the same group or individual is prohibited.
3. No restriction exists on the number of groups from an institute or organization, but common group participants are not allowed.
4. The groups or individuals must upload the obtained results in the specific format to the challenge website, which organizers will verify.
5. Automated scripts will calculate the scores for the proposed task and display the results on the leaderboard.
6. The group members or individuals can update the results on the leaderboard, with a limit of 5 per day per group, after the test data release and before the end of the competition.
7. To verify the results and ensure fair participation, the participants or top teams will be told to reproduce them at the event (or before) using the trained models. If the submitted results are mismatched, the participants will be disqualified, and their results will be removed from the leaderboard.

The winning teams are: Sahajeevi (Suman Paul Choudhury, Rohit Singh) and IAMGROOT (Sandeep Nagar). The baselines are created by Anandita Jamwal and Vibhu Gupta for task 1 and task 2, respectively.

5 Methodology

This section describes the baseline methodology, along with the approach of the winning teams, for both the tasks of the C4MTS challenge.

5.1 Task 1: Traffic Signs Detection

The task is to train a model to detect traffic signs using the provided bounding box annotations. The output expected for evaluating this task is the (*class name*, *x center*, *y center*, *width*, *height*, *prediction score*) for each frame. The *Mean Average Precision* (mAP) is used to evaluate the output.

Baseline Method Considering its light and efficient architecture, this approach uses the YOLOv8n [11] model. The challenge dataset is divided into a 90 : 10 train-validation ratio. The SGD optimizer is used to train the model for 100 epochs with a batch size of 16, and the prediction is made on the test data.

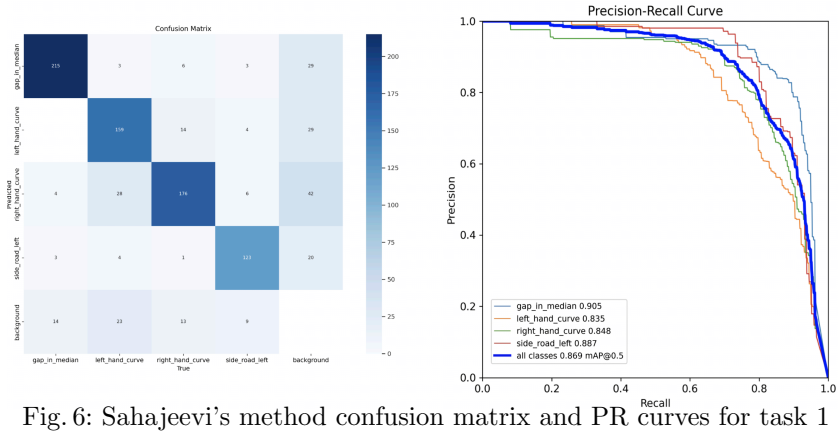


Fig. 6: Sahajeevi's method confusion matrix and PR curves for task 1

Sahajeevi's Method First, the following data augmentation methods [12] are applied: ShiftScale, random brightness-contrast, changing saturation level, and mosaic data augmentation [13]. The SoTA YOLOv8 [11] object detection model is trained on an NVIDIA A100 GPU, taking about 40 minutes to train 60 epochs. For post-processing and inference, SAHI [14] detects small traffic signs. Evaluation plots are given in Fig. 6.

IAMGROOT's Method The IMAMGROOT team applied the following data augmentation techniques from Roboflow: random rotation, and mosaic data augmentation, horizontal flip, and vertical flip, and generated 8000 extra samples from the task-1 dataset, resulting in total 10000 instances. It is observed that adding Dropout also improved the results. The small footprint of the traffic sign in the images suggests that this is a fine-grained detection task. Given the relatively limited samples in the dataset, small-to-medium models are considered to avoid possible overfitting. The task-1 dataset is split into train and validation, with an image size of 720×720. Existing SoTA models: YOLOv8n, YOLOv8s, YOLOv8m, and YOLO8l are fine-tuned using pre-trained weights, with the YOLOv8s model being optimal. The models are trained on the NVIDIA A100 for 80 epochs with the following hyper-parameters: $1e-4$ learning rate, $2e-4$ weight-decay, 0.982 momentum, 0.25 Dropout.

5.2 Task 2: Traffic Sign Scene Categorization

The aim of Task 2 is to train a model wherein model training happens using road scene images with in-painted traffic signs spread over four traffic sign categories: *left-hand-curve*, *right-hand-curve*, *gap-in-median*, and *side-road-left* for scene categorization. The evaluation for task 2 is done by comparing the predicted labels submitted by participants to the test labels of the 200 test images, with top-1 accuracy as the chosen metric to evaluate the models. The test set contains a proportionate combination of real and in-painted images to ensure that the model doesn't favor the in-painting artifacts and performs poorly on real scenes.

Baseline Method For the second task, the ResNet-18 [15] model is chosen. The following hyper-parameters are employed for the experiment: 20 batch size, 100 epochs, 0.01 learning rate, $1e-4$ weight decay, and momentum of 0.9. However, the following hyper-parameters were observed to be optimal: 2 batch size, 100 epochs, 0.008 learning rate, $1e-6$ weight decay, and a momentum of 0.9, with the SGD optimizer used throughout. The model with the optimal parameters is further tested on the CIFAR-10 dataset, resulting in an impressive 0.91 accuracy. All models are trained on Google Colab, following a reference code[2].

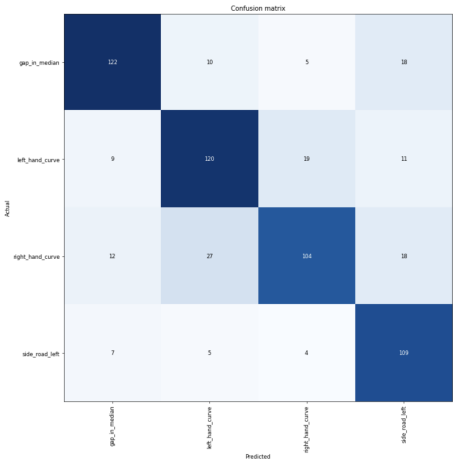


Fig. 7: Sahajeevi's method confusion matrix for task 2.

Sahajeevi's Method No data augmentations are performed, and the preferred model is the ResNet50 [15], which is fine-tuned on the challenge dataset for 10 epochs. The confusion matrix is shown in Fig. 7.

IAMGROOT's Method The IAMGROOT team improved the SoTA performance by conducting fine-tuning on the ResNet18 architecture for four-class

[2] ResNet-18 reference code link

classification. To further improve the model, the team utilized data augmentation techniques. For model training, the task-2 dataset was split in 90 : 10 train-validation ratio, and images were normalized using the following mean and standard deviation values: $(0.4914, 0.4822, 0.4465)$, $(0.2023, 0.1994, 0.2010)$ respectively. The model is initialized using pre-trained weights and fine-tuned for 60 epochs with the following hyperparameters: $8e-4$ learning rate, 0.9 momentum, $1e-6$ weight decay, and the SGD optimizer. The model is evaluated using the Cross-Entropy loss. Resuls are described in 2, and the code is made available[3].

Table 2: Task 1 and Task 2 results

Method	**mAP** (Task-1)	**Top-1 Accuracy** (Task-2)
Baseline	0.88	0.290
Sahajeevi	0.84	0.514
IAMGROOT	**0.90**	**0.605**

6 Results

The results are elaborated in Table 2. The IAMGROOT's method outperforms the baselines with a mAP of 0.90 in task 1, while Sahajeevi's method underperforms the baseline by 4.76%. This happens perhaps because the baseline is trained for 40 more epochs than Sahajeevi's method (refer Sec 5). For task 2, both teams significantly outperformed the baseline, with IAMGROOT being the best overall, with a top-1 accuracy of 0.605.

7 Session held during the Event

We thank IHub-Data, IIIT Hyderabad (https://ihub-data.iiit.ac.in/), for sponsoring the awards. We had a dedicated two-hour session at NCVPRIPG to highlight the importance of the problems related to missing objects and provide the winning teams with a platform to showcase their work. The session included the presentation of the challenge, award ceremony, three oral presentations, and a keynote talk on *Recognition of Modern Indian Signboards and License Plates* by Prof. Ganesh Ramakrishnan (IIT Bombay).

8 Conclusion

We conducted the C4MTS challenge on Categorizing Missing Traffic Signs to gather interest in an exciting and equally important task of improving road infrastructure. For this, the challenge was designed to contain two tasks: Traffic sign detection and Traffic scene categorization, focusing on four categories of

[3] github.com/Naagar/Missing_Traffic_Sign

traffic signs from the recent Missing Traffic Signs Video Dataset. Methodologies explored by the winning teams, who beat the baseline top-1 accuracy by a significant margin of 31.5% were also discussed. We encourage others to research and explore the publicly released Missing Traffic Signs Video Dataset.

References

1. Akshay Jadhav. Autonomous vehicle market by level of automation (level 3, level 4, and level 5) and component (hardware, software, and service) and application (civil, robo taxi, self-driving bus, ride share, self-driving truck, and ride hail)-global opportunity analysis and industry forecast, 2019-2026. *Allied Market Research*, 2018.
2. World Health Organization. Road traffic accidents, 2022.
3. Dorde Petrovic, Radomir Mijailovic, and Dalibor Pesic. Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. *Transportation research procedia*, 45:161–168, 2020.
4. Varun Gupta, Anbumani Subramanian, CV Jawahar, and Rohit Saluja. Cue-can: Cue driven contextual attention for identifying missing traffic signs on unconstrained roads. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
5. Eugene Chian, Weili Fang, Yang Miang Goh, and Jing Tian. Computer vision approaches for detecting missing barricades. *Automation in Construction*, 131:103862, 2021.
6. Jin Sun and David W Jacobs. Seeing What Is Not There: Learning Context to Determine Where Objects Are Missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2017.
7. Varun Gupta. Dashcam Anonymizer. github.com/varungupta31/dashcam_anonymizer, 8 2023.
8. Jui-Ting Chien, Chia-Jung Chou, Ding-Jie Chen, and Hwann-Tzong Chen. Detecting Nonexistent Pedestrians. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 182–189, 2017.
9. Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2149–2159, January 2022.
10. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
11. Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023.
12. Alexander B. Jung. imgaug. https://github.com/aleju/imgaug, 2018. [Online; accessed 30-Oct-2018].
13. Alexey Bochkovskiy, Chien-Yao Wang, and Hong-yuan Liao. Yolov4: Optimal speed and accuracy of object detection, 04 2020.
14. Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, oct 2022.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.