IMPROVED TRAFFIC SIGN DETECTION IN VIDEOS THROUGH REASONING EFFECTIVE ROI PROPOSALS

Yanting Zhang**, Yonggang Qi***, Jie Yang***, and Jenq-Neng Hwang*†

*Beijing University of Posts and Telecommunications, China; †University of Washington, Seattle, USA
{zhangyt, qiyg, janeyang}@bupt.edu.cn; hwang@uw.edu

ABSTRACT

Traffic sign detection is an important task in assisted safety and autonomous driving. It is important to continuously detect the traffic signs emerged on the road. Currently, most object detection methods make independent detections based on single images. When we apply these methods directly to a video clip to detect traffic signs without taking into account temporal correlations among adjacent frames, missed detections or incorrect detections can frequently occur due to motion blur, size change, partial occlusion, and/or bad pose. In this paper, we fully exploit the temporal consistency of traffic sign detection in videos. More specifically, we incorporate information of adjacent frames with high confidence scores to enhance the discovery of potential objects in the missed or incorrect detected frames by "recovering" the missed RoI proposals or by "improving" the incorrect RoI proposals with low confidence scores. Our method can be regarded as a "detection-by-tracking" strategy, which results in a more robust detection performance in videos.

***Index Terms*—** Traffic sign, video object detection, RoI proposals, detection-by-tracking

1. INTRODUCTION

In recent years, both academic and industrial communities pay great attention to autonomous driving. Traffic signs are important components in driving scenarios. Not only can traffic signs provide reliable navigation, but also they can be a good reference to facilitate other applications in autonomous driving as they always have standard size, pattern, and shape. For example, they can be utilized to create reliable correspondence among frames to carry out robust local bundle adjustment in simultaneous localization and mapping (SLAM), contributing to a better vehicle self-localization [1, 2]. It is thus critical for an autonomous driving system to reliably detect and track the traffic signs in video sequences.

Previously, traffic signs are usually detected and classified through color-based as well as shape-based methods [3]. In recent years, with the major advances of convolutional neural networks (CNNs) based detectors [4, 5, 6], we have observed a large improvement in detection accuracy and efficiency [7]. However, most traffic sign detections are mainly carried out based on single still images [7, 8], in spite of the fact that a moving dash camera captures the environment continuously in a driving scenario, which should mandate a more effective way of detecting the traffic signs in a video. If we just simply apply the independent object detection methods on images, missed detections (i.e., no bounding boxes are identified on the objects) or incorrect detections (e.g., incorrect classifications, false positive detections, incorrect locations and/or sizes of bounding boxes) can be frequently observed due to motion blur, partial occlusion, scale change, and bad viewing perspective [9, 10]. Some examples of single-image missed/incorrect detection results in video sequences are shown in Figure 1. The detection mechanism in this way is problematic since we totally ignore the temporal correspondence among image frames within a video sequence.

In this paper, we focus on improving the traffic sign detection performance in autonomous driving scenarios. To detect traffic signs effectively and robustly in videos, we exploit the temporal consistency among image frames, i.e., the highly correlated location and appearance information of the same traffic sign among consecutive image frames. Our proposed method is based on the two-stage object detector, Faster R-CNN [4], where the Region Proposal Network (RPN) identifies a bunch of object proposals and the multi-task classification and bounding box regression are then carried out on the pooled features within these region of interests (RoIs). When the traffic sign is in bad condition, the pooled feature cannot well represent its property, resulting in missed detections or incorrect bounding box regression. Besides, Softmax used in classification can also mislead the object to a wrong classification result because traffic signs in the same big category are pretty similar when the signs are relatively small. If we can leverage the good detection results with high confidence scores in adjacent frames, the detection performance of the whole sequence can thus be improved to a large extent, i.e., either by recovering the missed detections or improving the incorrect detections.

1) We explore a detection-by-tracking mechanism for traffic sign detections in videos to benefit the autonomous driving in recovering or improving detections of objects more effectively and robustly.

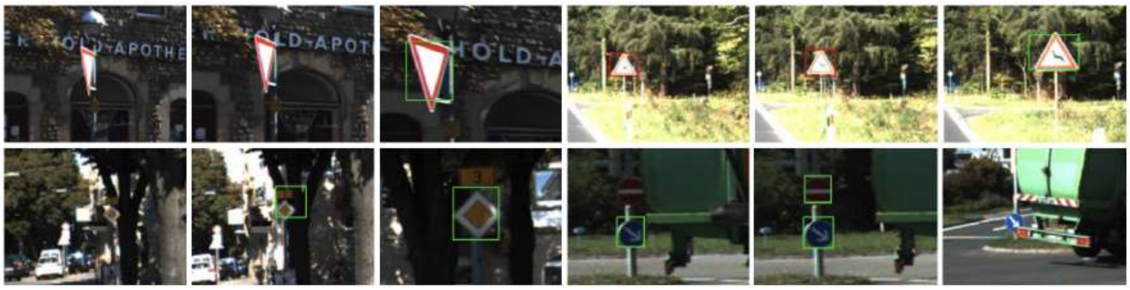

Fig. 1. Examples of traffic sign detections from a single-image based detector. Red boxes denote incorrect classes and green boxes denote correct classes, note that many missed detections and incorrect detections with wrong sizes and locations exist.

2) The temporal correlation information of the same traffic sign among video frames is fully exploited based on our shortest path search over all the candidate object proposals. The classification and regression performance of promising object proposals can be greatly enhanced using consecutive frames' location/size and pooled feature information.

3) Our method can solve the detection bottleneck of most single-image based models, where the detection performance may be limited due to practical reasons, such as limited training samples.

The rest of the paper is organized as follows. In Section 2 we survey the related works. Our proposed system is introduced in Section 3, aiming at solving the key issues raised in the baseline method of single-image based video object detections. Experimental results and discussions are given in Section 4, followed by the conclusion in Section 5.

2. RELATED WORKS

In autonomous driving, it is essential to have a continuous and reliable perception of the outside environment. Thanks to the great advances of deep convolutional neural networks (CNNs), object detection [4, 5, 6] in images gradually becomes an easy task and provides the driverless car with thorough information. Traffic sign detection also experiences a considerable improvement when employing the deep learning based techniques [7]. Widely used object detection methods can be divided into single-stage and multi-stage detectors. YOLO [5] and SSD [6] are representatives of single-stage object detectors, which identify and localize objects directly over a dense sampling of possible locations. On the other hand, Faster R-CNN [4], also known as a region-based method, divides the object detection process into two stages: 1) region proposals are first generated through selective search or a regional proposal network (RPN); 2) the multi-task classification and bounding box regression are then carried out on the region candidates. In general, region-based methods, which pass the proposal candidates with a higher object-containing probability to the subsequent object classification and bounding box regression tasks, can produce better detection performance compared with the single-stage methods [11].

However, there is an acknowledged problem in single-image based object detection, i.e., the detection performance only independently relying on a single image is not reliable when dealing with different image quality within a sequence due to motion blur, video defocus, partial occlusion or bad pose. People gradually pay attention to leverage the temporal cues for object detection in video sequences [9, 10, 12, 13]. Zhu et al. [9] propose a feature aggregation along motion path guided by an optical flow scheme to improve the feature quality. Similarly, Wang et al. [10] propose a fully motion-aware network to jointly calibrate the object features on pixel-level and instance-level. They both operate on the extracted features to pursue better feature representations. However, flow estimation is very time consuming, and the aggregated features cannot guarantee accurate and robust results with missed and/or incorrect detections still occurring. Kang et al. [12] propose a tubelet proposal network to generate spatially associated bounding boxes across time, further processed by a Long Short-term Memory (LSTM) recurrent network. The method cannot deal with large motion since the initialized proposals are aligned at the same position in consecutive frames. Feichtenhofer et al. [13] tune a frame-based object detection and across-frame track regression network to improve the simultaneous detection and tracking performance. The aforementioned methods all perform end-to-end training, i.e., adequate training samples are required. From a post-processing perspective, Han et al. [14] propose to boost potential detections with lower scores by using detected objects with higher confidence scores from nearby frames. However, the objects are purely selected based on Intersection-over-Union (IoU) among frames to form a sequence without taking into account the embedded features of the proposal objects. These video object detection methods are all performed based on the dataset of ImageNet VID dataset [15] where the camera movement is moderate. Though the ideas are motivating and inspiring, they are not suitable in the autonomous driving scenarios due to extraordinary viewing changes from a dash camera mounted on a moving car.

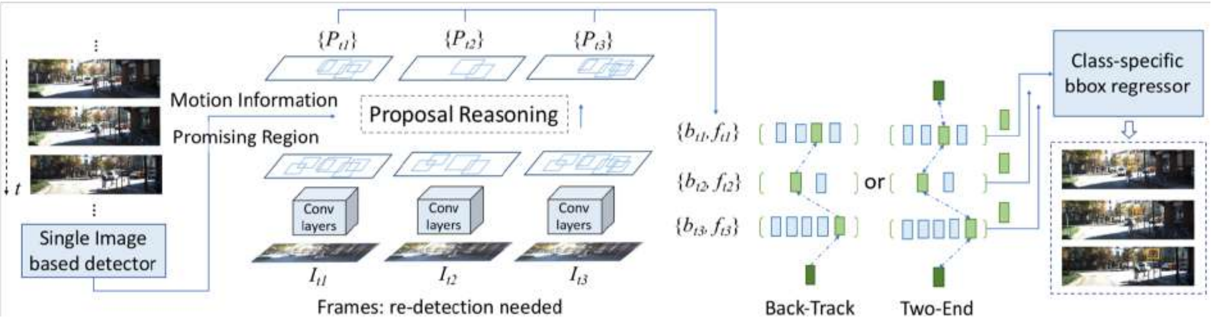

Fig. 2. The framework of our proposed algorithm.

The most similar work with our approach is [16], where multi-object tracking is explored purely based on an object detector. We both pay attention to object proposals. Differently, Bergmann et al. [16] use the detected object' location of the current frame to initialize the next frame's object proposal, following a tracking-by-detection workflow. In our paper, to mitigate the influences from large motion, we adopt a different procedure for proposal selection, both considering the motion cues and feature similarity among adjacent frames, which can lead to a more robust and accurate detection result.

3. PROPOSED DETECTION-BY-TRACKING SCHEME

In this section, we will address in details our innovative approach for traffic sign detection in videos, as shown in Figure 2.

We assume there are T image frames in a video sequence, $\{I_1, I_2, ..., I_T\}$, and there are K distinct classes of the traffic signs. Through our proposed method, we want to detect the traffic signs in every image frame I_t within the video sequence, $\{b_t^i, c_t^i | t \in T, c_t \in K\}$, where $b_t^i = (x_t, y_t, w_t, h_t)$ and c_t^i denote the bounding box and the class label of the i-th object in the t-th frame. As is shown in Figure 2, the whole framework consists of the following 5 steps.

1) Traffic sign detection based on single-image object detection. We use a multi-task loss L to train a Faster R-CNN detector. Let L_{cls} denote the cross-entropy loss and L_{reg} be the smooth L_1 loss, thus, L is defined as,

$$L(p, q) = L_{cls}(p, p^*) + \lambda p^* L_{reg}(q, \phi(f)), \tag{1}$$

where the hyper-parameter λ balances the two losses; p represents the class probability for each RoI, and p^* corresponds to the ground-truth label. q represents the regression target, while $\phi(f)$ is the predicted bounding-box regression offset, which takes the pooled feature as input.

We feed each image frame in $\{I_1, I_2, ..., I_T\}$ into the single-image object detector, which gives a preliminary detection result $\{B_t, C_t\}$ for I_t, where $B_t = \{b_t^1, b_t^2, ...\}$ and $C_t = \{c_t^1, c_t^2, ...\}$. This step only regards the video sequence as a collection of independent images and ignores the temporal consistency of objects in consecutive frames. It inevitably leads to fluctuations in detection performance, e.g., some frames with traffic signs have either no traffic sign detected (missed detections) or incorrect detections with wrong classes or deviated locations/sizes of bounding boxes.

2) Proposal identification in between two detected signs (Two-End). Through Step 1, we observe some missed detections or incorrect detections. Because the same traffic sign's embedded features should be similar among adjacent frames and its location should change gradually, which is consistent with the car moving pattern, the temporal consistency among frames should be exploited. For the i-th traffic sign, we assume that they have been both detected successfully in frames $I_{t'}$ and $I_{t''} (t'' > t')$ with high confidence scores on the same traffic sign class c^i and missed or incorrect detections occurred in between. Leveraging the "track" information, we use the detected boxes of the i-th object, which are class labeled and bounding box regressed from the 2^{nd} stage of the Faster R-CNN, i.e., $b_{t'}^i$ and $b_{t''}^i$, to generate a promising region R^i for either the missed or incorrectly detected signs between the two high-confidence detected frames of $I_{t'}$ and $I_{t''}$. Here, the promising region R^i is defined as the smallest rectangle which covers both $b_{t'}^i$ and $b_{t''}^i$. Within R^i, for every missed or incorrectly detected frame I_t we can keep a set of candidate object proposals $\{P_t\}$, as provided from the Faster R-CNN.

3) Backtracking proposals from a detected sign (Back-Track). Through Step 1, we may only detect the traffic signs starting from frame $I_{t''}$ with high confidence score, with missed or incorrect detections occurring before frame $I_{t''}$. Based on the moving trajectory of the detected traffic signs from $I_{t''}$, we can infer the promising regions for either the missed or incorrectly detected signs before $I_{t''}$, which are normally smaller and difficult to be correctly detected. Through this step, for every missed or incorrectly detected frame I_t (couple frames before $I_{t''}$) we can keep a set of candidate object proposals $\{P_t\}$, as provided from the Faster R-CNN.

4) Shortest path search based on RoI aligned features of regressed boxes. We assume the RoI aligned feature related to the proposal box $b_t^i \in P_t$ is f_t^i for the i-th traffic sign in

frame I_t. A shortest path can be created for traffic sign between $I_{t'}$ and $I_{t''}$ in a "Two-End" mode, with $\{b^i_{t'}, f^i_{t'}\}$ and $\{b^i_{t''}, f^i_{t''}\}$ being served as the starting and ending nodes (in case of Step 3, there will be no specific starting node, while a shortest path can still be generated as a "Back-Track" mode shown in Figure 2). The proposals $\{P_t | t' < t < t''\}$ in the promising region R^i in image frames $\{I_t | t' < t < t''\}$ can serve as intermedia nodes. Only proposals in adjacent frames have edges, whose edge costs are defined as Euclidean distance of RoI aligned features extracted from the two proposals located in corresponding adjacent frames. After constructing the graph, we perform the Dijkstra algorithm [17], a shortest path algorithm, to return a path from $\{b^i_{t'}, f^i_{t'}\}$ to $\{b^i_{t''}, f^i_{t''}\}$. The objective function can be defined as Equation 2, where $d()$ is the distance measure. The features of selected intermedia nodes in $\{P_t\}$ are considered to be the most similar and consistent with the starting and ending nodes, i.e., previous and future detections for Step 2. On the other hand, for Step 3, we find the prior missed detections which are most consistent with the later detection $b^i_{t''}$.

$$\underset{j_{t'},\dots,j_t,\dots,j_{t''}}{argmin} \sum_{t'}^{t''} d(f_t(j_t), f_{t+1}(j_{t+1})), \qquad (2)$$
$$s.t., t' \le t < t+1 \le t'', j_t \in \{P_t\}.$$

5) Class-specific bounding box regression. Either using "Two-End" (Step 3) or "Back-Track" (Step 4), we keep the proposals which are expected to contain the specific traffic sign in missed detection or incorrect detection frames. For each selected proposal box (x, y, w, h) with RoI feature f, we carry out a class-specific bounding box regression. To recover from the predicted parameterized coordinates output $(\phi_x(f), \phi_y(f), \phi_w(f), \phi_h(f))$, we use Equation 3 to get the final regressed bounding box $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$.

$$\begin{aligned} &\hat{x} = w\phi_x(f) + x, \hat{y} = y\phi_y(f), \\ &\hat{w} = we^{(\phi_w(f))}, \hat{h} = he^{(\phi_h(f))}. \end{aligned} \qquad (3)$$

In all, compared with single-image detection methods which offers the detections based solely on the classification confidence scores independently, we make use of the detection results of frames with high confidence scores, along with the temporal consistency of refined RoI aligned features, to help discover the potential objects in missed or incorrect detection frames through careful associations of the proposals.

4. EXPERIMENTS

4.1. Dataset

To train a basic single-image based traffic sign detector, we use the German traffic sign benchmark (GTSDB) [8] because it provides detailed ground truth annotations. The images are collected from different scenarios, which can help our detector to learn and generalize better in the wild. GTSDB and KITTI [18] are both collected in Germany, thus they contains the same traffic sign types. We choose several sequences, 0005, 0014, 0015, 0029 and 0084, from the KITTI raw dataset to evaluate our proposed approach about video-based traffic sign detection. The sequences, in total 1578 frames, are chosen because traffic signs are frequently observed. We have labeled the traffic signs with a tight bounding box as the ground truth in KITTI for performance comparison.

4.2. Experiment Setup

To train a baseline single-image based traffic sign detector, we have implemented the Faster R-CNN based on [19]. The shared convolutional layers are initialized by a pre-trained model for ImageNet classification (ResNet R-50-C4). A learning rate of 0.0025 is used for the first 12K iterations, and is then decreased by 0.1 each time for another 4K and further 2K iterations, until it stops at 18K iterations. The experiments are carried out on an NVIDIA Quadro GV100 32GB GPU.

4.3. Results

The single image based Faster R-CNN traffic sign detector is not perfect enough due to limited training samples and challenging wild images. From our observations, the detection results for the same traffic sign in a video through the single-image based detector is fluctuating, forming a series of tracklets separated by the missed/incorrect detections. In general, the mistakes are likely to occur when the sign is far away or when the patterns of signs are pretty similar. Following the steps described in Section 3, we use the information from neighboring frames to boost the discovery of signs in the missed and incorrect frames.

Due to limited number of training video clips with traffic signs, it is tough to employ the video object detection methods like [9, 10, 12], though these methods are not proper in autonomous driving scenarios with large motion changes. For a fair comparison, we implement the method proposed in [16], entitled Tractor-based, where the regression of the Faster R-CNN object detector aligns the existing bounding boxes in frame I_{t-1} to the object's new position at frame I_t. We set a tolerant threshold of 0.001 for accepting the new detection, which will be used as an initialization of object position in frame I_{t+1}. The tracking will be interrupted once the classification confidence score is not supportive, in order to prevent potential false positives.

We use Average Precision (AP) and Average Recall (AR), averaged over all traffic sign categories with the 0.5 IoU threshold, to describe the detection performance. Table 1 shows the results, where the AP is 0.628 and AR is 0.626 for the basic Faster R-CNN detector on the chosen KITTI sequences. Taking into account the temporal information, both Tractor-based and our proposed Shortest Path (SP) based methods can achieve a better performance. The AP can even

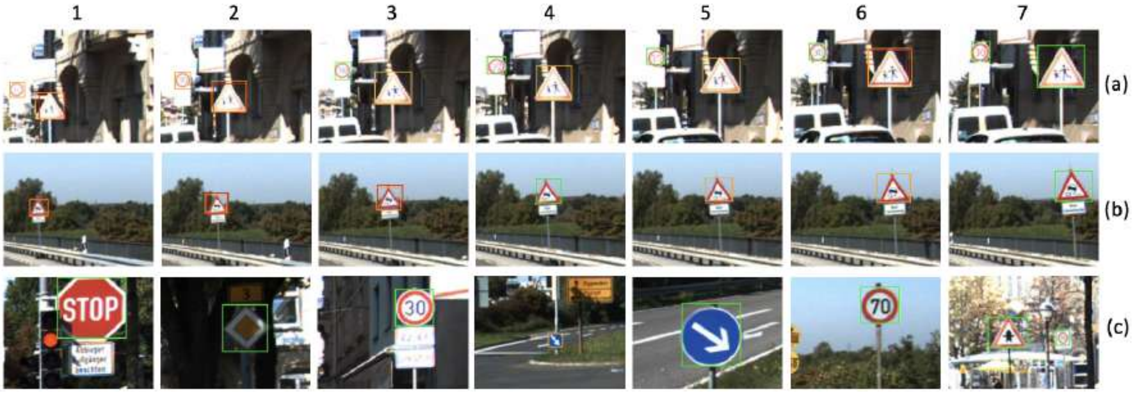

Fig. 3. Qualitative performance of recovered or improved bounding boxes from originally missed/incorrect detections. Red boxes denote incorrect detections either due to wrong classification or wrong location/size of bounding boxes, green boxes denote the correct detections with high confidence, and orange boxes denote the missed or incorrect detections recovered by our proposed method with correct class labels and accurate localizations.

Table 1. The performance of different methods.

Method	AP@.50	AR@.50
Faster R-CNN	0.628	0.626
Tractor-based	0.640	0.647
Ours (SP)	**0.733**	**0.748**

be improved to 0.733 and AR reaches 0.748 for our proposed algorithm. The performance of the Tractor-based method is significantly inferior to that of ours, it is because the moving camera in driving scenarios brings about huge differences among frames, resulting in low-quality box initializations.

Figures 3(a) and (b) qualitatively show recovered frames of detection results of our proposed algorithm in two example sequences and Figure 3(c) shows detection-by-tracking results of various traffic signs in our dataset. In Figure 3(a), the triangular traffic sign is only detected with high confidence on the last frame, and the remaining three missed frames (frames 3,4,5) before it and three incorrect detections of wrong sizes (frames 1,2,6, with red boxes) are all recovered or corrected as shown with orange bounding boxes. Similarly, the five detected circular speed limit signs with green bounding boxes are used to successfully recover the two missed signs with orange boxes before it. In Figure 3(b), two missed detected signs (frames 5,6) are recovered (with orange boxes) from two detected high-confidence green boxes (frames 7,4), while three incorrect detections (frames 1,2,3) are remedied through backtracking based on frame 4.

4.4. Shortest Path vs. Maximum Score

To quantitatively evaluate the effectiveness of our proposed method, we also compare our method with another procedure regarding proposal selection, i.e., maximal confidence score based method (MS). Regressed proposal with the maximal confidence score of the same sign category is chosen in the promising region for MS method, whose objective function can be defined as Equation 4, where $s_c()$ is the confidence score of the same sign category c.

$$\underset{j_t}{argmax} \sum_{t'}^{t''} s_c(f_t(j_t)), s.t., t' < t < t'', j_t \in \{P_t\}. \quad (4)$$

Three metrics are used to describe the discrepancy between predictions and ground truths following the protocols in [12]. Mean absolute pixel difference (MAD), mean relative pixel difference (MRD), and mean IoU between predicted boxes and target boxes are calculated to evaluate the bounding box detection quality. We give a comparison between SP and MS in Table 2, where we show the performance of two different recovering schemes, i.e., of the "Two-End" (with starting and ending high confidence detections) and "Back-Track" (from the last frame to recover prior missed detections).

Table 2. The performance of MS and SP methods.

	Overall			Two-End			Back-Track		
	MAD	MRD	IoU	MAD	MRD	IoU	MAD	MRD	IoU
MS	2.31	0.104	0.681	2.65	0.119	0.634	1.81	0.082	0.751
SP	2.06	0.093	0.706	2.31	0.103	0.667	1.68	0.079	0.765

We can see that both short-path based search and max-score based search can find the objects with only a slight offset with ground truth labels. SP has a better performance in locating objects, though it may hold a lower confidence score returned by the original Faster R-CNN detector. While MS ignores the temporal RoI feature consistency, SP leverages the similarity of refined RoI pooled feature in adjacent frames and thus contributes to a better performance.

5. CONCLUSION

In this work, we propose a framework based on a two-stage object detection method for more robust traffic sign detection in videos. We focus on discovering the best proposal generated by RPN of the trained Faster R-CNN detector based on the generic single-image object detection result. Simulations show that our framework can achieve a consistent performance improvement over single-image based detections. The proposed detection-by-tracking scheme correlates the temporal information throughout frames in a video and can provide more robust and reliable detection results. Besides, our method mainly focuses on the inference phase without any modification in the training phase, therefore our framework can be easily applied to other single-image region-based detection networks to improve the performance in video object detection.

6. ACKNOWLEDGMENT

This work is supported partially by the National Natural Science Foundation of China (61671078, 61701031, 61601042), Funds of Beijing Laboratory of Advanced Information Networks of BUPT, Funds of Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, the National Key Research and Development Program of China(2017YFB0802701), the Fundamental Research Funds for the Central Universities (2018XKJC04) and 111 Project of China (B08004, B17007).

7. REFERENCES

[1] Xiaozhi Qu, Bahman Soheilian, and Nicolas Paparoditis, "Vehicle localization using mono-camera and geo-referenced traffic signs," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 605–610.

[2] Yanting Zhang, Jie Yang, Haotian Zhang, and Jenq-Neng Hwang, "Bundle adjustment for monocular visual odometry based on detected traffic sign features," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

[3] Andrzej Ruta, Yongmin Li, and Xiaohui Liu, "Real-time traffic sign recognition from video by class-specific discriminative features," *Pattern Recognition*, vol. 43, no. 1, pp. 416–430, 2010.

[4] 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*, 2015, pp. 91–99.

[5] 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*, 2016, pp. 779–788.

[6] 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*. Springer, 2016, pp. 21–37.

[7] Yanting Zhang, Ziheng Wang, Yonggang Qi, Jun Liu, and Jie Yang, "Ctsd: A dataset for traffic sign recognition in complex real-world images," in *2018 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2018, pp. 1–4.

[8] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *The 2013 international joint conference on neural networks (IJCNN)*. IEEE, 2013, pp. 1–8.

[9] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 408–417.

[10] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng, "Fully motion-aware network for video object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 542–557.

[11] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele, "What makes for effective detection proposals?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 814–830, 2016.

[12] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang, "Object detection in videos with tubelet proposal networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 727–735.

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Detect to track and track to detect," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3038–3046.

[14] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe, "Tracking without bells and whistles," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 941–951.

[17] Donald B Johnson, "A note on dijkstra's shortest path algorithm," *Journal of the ACM (JACM)*, vol. 20, no. 3, pp. 385–388, 1973.

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

[19] Francisco Massa and Ross Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," `https://github.com/facebookresearch/maskrcnn-benchmark`, 2018.