CueCAn: Cue-driven Contextual Attention for Identifying Missing Traffic Signs on Unconstrained Roads

Varun Gupta⌉, Anbumani Subramanian†, C.V. Jawahar†, Rohit Saluja‡

https://github.com/iHubData-Mobility/public-CueCAn

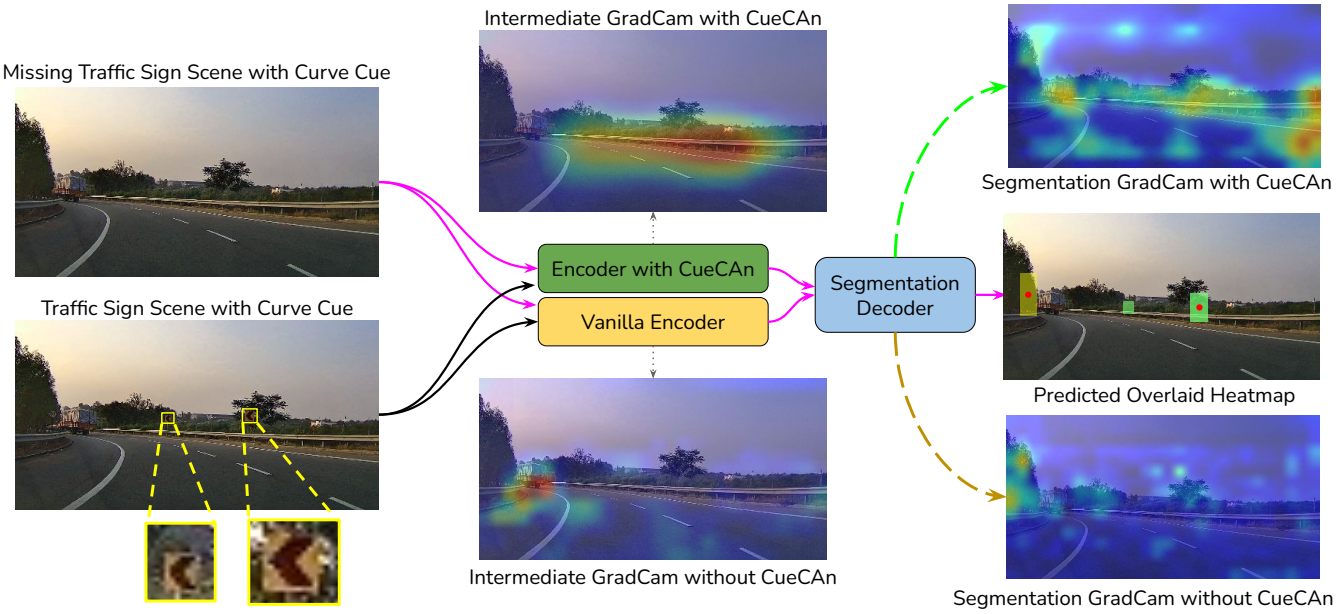

Fig. 1: **Left:** Scenes with real and inpainted traffic signs (chevron-left). **Middle:** Intermediary GradCam visualizations of the cue classifier (encoder) with and without CueCAn. **Right:** Segmentation model with CueCAn-based encoder detects missing signs (green masks overlayed over the scene on the right for CueCAn and yellow mask by the baseline) on the scene without signs (follow pink arrows) by effectively attending to the context cues, compared to weak attention without CueCAn. Segementation GradCams are obtained from the centroid of the predicted sign (red dot).

***Abstract*— Unconstrained Asian streets often involve poor infrastructure, affecting overall road safety. Missing traffic signs are a regular part of such streets. Missing or non-existing object detection has been studied for locating missing curbs and estimating reasonable regions for pedestrians on road scene images. Such methods involve analyzing task-specific single object cues. In this paper, we present the first and most challenging video dataset for missing objects, with multiple types of traffic signs for which the cues are visible without the signs in the scenes. We refer to it as the Missing Traffic Signs Video Dataset (MTSVD). MTSVD is challenging compared to the previous works in two aspects: i) The traffic signs are generally not present in the vicinity of their cues, ii) The traffic signs' cues are diverse and unique. Also, MTSVD is the first publicly available missing object dataset. To train the models for identifying missing signs, we complement our dataset with 10K traffic sign tracks, with 40% of the traffic signs having cues visible in the scenes. To solve the problem of identifying missing traffic signs, we propose novel Cue-driven Contextual Attention units (CueCAn), which we incorporate into our model's encoder. We first train the encoder to classify the presence or absence of traffic sign cues and then train the entire segmentation model end-to-end to localize missing traffic signs. Quantitative and qualitative analysis shows that CueCAn significantly improves the performance of base models. Code, data, and models will be released. Refer GitHub link and supplementary for the demo.**

⌉†‡ The authors are with Center for Visual Information Technology (CVIT) Lab, IIIT Hyderabad, India. ⌉ varungupta.iiith@gmail.com, † {anbumani,jawahar}@iiit.ac.in, ‡rohit.saluja@research.iiit.ac.in

I. INTRODUCTION

Road accidents affect millions of lives due to increasing population and vehicle density. Incorporating Advanced Driving Assistant Systems (ADAS) in commercial vehicles is gaining momentum. However, these systems still have a long way to go in terms of offering absolute road-safety [1]. In a recent road safety report, it is observed that 12.6% of accidents caused by the driver errors are due to traffic sign violations, which establishes the importance and need of traffic signs in a regulated manner [2]. Autonomous Vehicles (AV) require a consistent and well-maintained infrastructure to explore their full potential and to deliver the safety they promise to offer. Traffic signs are an important aspect of road infrastructure, as they contain essential information about what is coming ahead, which ADAS systems like Mobileye use as supplementary information for scene understanding. Failure to robustly perceive and process the road scene has led to multiple fatal crashes involving commercially deployed AVs as well [3]. Missing traffic signs are frequent on the unconstrained roads of many Asian countries; few samples of such frames from our Missing Traffic Signs Video Dataset (MTSVD), containing context cues but no signs are present in Fig. 2. Identifying regions with missing traffic signs in a manual manner requires significant effort. Prior missing object datasets either had the cue in their immediate vicinity or had a consistent relationship between the missing

Manuscript 2450 submitted to 2023 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2022.

Fig. 2: Sample scenes from MTSVD exhibiting missing traffic signs (with cues). **Top (left to right)**: side-road-left, cross-roads, go-slow and bus-bay. **Bottom (left to right)**: right-hand curve, gap-in-median and left-hand curve.

TABLE I: Overview of Missing Object Datasets.

Dataset	Frames with Missing Objects	Public Access	Context Variety
Tohme [5], [6]	1086	×	×
Pedestrians Dataset [4]	<475	×	×
Missing Barricades [7]	853	×	×
MTSVD (ours)	**135K (2K Videos)**	✓	✓

object and the corresponding cue [4], [5]. However, the MTSVD contains multiple cue contexts and complementary cue-object relationships, making MTSVD the most challenging and diverse, publicly accessible missing objects dataset. Based on the observation that context clues add a certain discontinuity in the scene, we also propose CueCAn to exploit the nature of context cues. The CueCAn works on the intuition of identifying traffic sign cues by erasing/inpainting them (or filling the cue regions with context) in feature space and then taking the difference between inpainted and the original features, which helps in highlighting the discontinuous cue patterns. The proposed pipeline, as illustrated in Fig. 1, highlights the efficacy of CueCAn in making the model attend to context features for both encoding and decoding tasks of classifying the absence/presence of cues and localizing missing signs, which are lacking in the vanilla approach, without CueCAn. Our contributions are as follows:

- We introduce the first publicly accessible video dataset for missing objects, the Missing Traffic Sign Video Dataset (MTSVD), spread across $10K$ video tracks for 101 traffic signs categories, with $2K$ missing sign video clips containing 20 types of traffic sign cues.
- We propose the Cue driven Context Attention Unit, *CueCAn*, a cue-driven approach for detecting missing traffic signs on unconstrained and traffic-dense roads.

II. Related Work

Missing Object Localization and Datasets: In the last decade, deep learning has driven computer vision approaches to be reasonably accurate for the tasks such as object detection and image segmentation. However, the task of locating missing objects is currently studied in very few works, but the existing ones highlight its potential. [4], [5], [7]. Humans consider multiple aspects to determine missing objects, the most prominent of which is cue understanding [8]–[10]. Sun et al. [5] aim to improve city accessibility for people with mobility disabilities by detecting regions with missing curbs. For this, a Siamese Fully Connected network (SFC) learns the contextual cue classifier for curbs from the image and object-masked image pairs, along with a curb localizer. The model identifies the region as a missing curb if an image contains curb context (cue) but no curb. Chian et al. [7] with an attempt to mitigate fall-from-height injuries, annotate the missing barricade regions, and perform object detection. Chien et al. [4] predict the regions where pedestrians could be placed in street scenes using the Fully Convolutional Network (FCN) based on the VGG encoder. Grabner et al. [11] processed context in video tracking tasks using supporters, which help estimate the target object locations using the Hough Transform.

An overview of existing missing object datasets is given in Table I. Given the lack of a large-scale and complex dataset specifically created for missing objects, prior missing object detection works either using existing datasets (generally curated for segmentation and object detection tasks) or collecting and sampling their own datasets lacking in scale and variety. Sun et al. [5] employ the TOHME dataset, a collection of 1086 street ramp images sourced partly from Google Street View and crowdsourcing [6]. Chien et al. [4] use the CityScapes dataset [12]. However, more than 84% of the CityScapes images contain fewer than 5% pedestrian pixels, making most frames unsuitable for the task [4]. Chian et al. [7] collect 853 images of barricades captured from high-rise construction sites using a crane-mounted camera at varying elevations. The datasets mentioned above do not involve diversity in the object cues. However, in the case of traffic signs, multiple cues exist, i.e., the cue for *left-hand-curve* is different from that of *right-hand-curve*, and the cue for *bus-bay* is much different from that of *pedestrian-crossing*, etc.

Traffic Sign Datasets: Several traffic sign datasets exist captured in different regions over the globe [13]–[16], [18].

Manuscript 2450 submitted to 2023 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2022.

TABLE II: Overview of Various Traffic Sign Datasets. *Tracks* refer to the video tracks with signs in a sequence, and *Missing signs* refer to intervals where cue-context exist, but the traffic sign does not. ‡ created using video data. † 52K signs fully annotated and 48K partially annotated [13]

Dataset	Capture Resolution	Frames (Tracks)	Night +All-weather	Missing Signs
DITS‡ [14]	1280×720	478	×	×
TT100K [15]	2048×2048	$26K$	×	×
GTSDS‡ [16]	1360×800	1206	×	×
BTSD [17]	1280×720	8851	×	×
MTSD [13]	2048×1152	$100K^{\dagger}$	✓	×
MTSVD‡ (ours)	**2560×1440**	**400K (10K)**	✓	✓

Fig. 3: Annotated frames from MTSVD exhibiting frames having signs with their relevant cues. **Top** (left to right): a roundabout with pedestrian-crossing, bus-bay with speed-limit. **Bottom** (left to right): u-turn-right with pedestrian-crossing (night scene), go-slow and Signboard (rainy weather).

Traffic signs spread across different countries though sharing a similar context and purpose, have variations in appearance, promoting the compilation of regional datasets. Table II briefly summarize different traffic sign datasets. Existing datasets focus on classification and segmentation tasks and leave out the missing signs data, resulting in the absence of a holistic approach toward road safety in unconstrained environments. The incentive behind collecting the MTSVD is further promoting road safety on unconstrained roads. Road safety depends on complex parameters such as the quality of existing road infrastructure, violation of traffic rules by other vulnerable road users (VRUs), including pedestrians (e.g., jaywalkers), and the driving behavior of road agents.

Context in Computer Vision: Elaborate studies indicate how humans effectively perceive their surroundings via contextual clues [8]–[10]. Santosh et al. [19] elaborate on the role of context in human-scene perception, classify context into multiple categories, and use context information to improve object detection. Multiple vision approaches integrate context and object features, like Conditional Random Fields, to improve object categorization and the Deformable Parts Model with surrounding context information to improve detection and segmentation tasks [10], [20]. However, given the close dependency between object features and their corresponding context, these approaches are unsuitable for identifying missing traffic signs, where the object remains absent from the scene.

A. *Deep Neural Networks and Attention Mechanism*

Deep neural networks have gained immense popularity in multiple computer vision tasks covering classification, semantic segmentation, and object detection. Fully Convolutional Neural (FCN) networks use skip connections between the up and the downsampling path, i.e., layers bypassing at least one intermediate layer in the network to effectively capture contextual and spatial information from the features [21]. FCN has been utilised for identifying missing pedestrians by Chien et al. [4], and the same has been used in this study for the segmentation task as well. For the segmentation tasks, the output matching the input size is generated with pixel-level class mappings. As we provide a plethora of data to neural networks, attention modules make the networks attend to specific input parts to improve the model's performance. Initially, Bahdanau et al. [22] introduced attention modules, which find applications in many AI applications, including Computer Vision, Natural Language Processing, Speech processing, etc. Liu et al. [23] propose a contextual attention unit considering global and local object features for obtaining the saliency maps. However, local information in the spatial vicinity of the signs is irrelevant for identifying missing traffic signs, as they are predominantly at a location away from their cue. Further, global information is not essential for our task as the cues for each traffic sign are at specific locations in the scene. Han et al. [24] provide a solution for detecting water puddles using the Reflection Attention Unit (RAU) on the basis of puddle surfaces containing reflection from the regions away from them (e.g., sky), which is similar to our situation (traffic signs away from their cue). However, RAU models search for similar corresponding patches in the continuous image space. In contrast, our task requires modeling traffic sign cues as a discontinuity in the image space, e.g., speed breakers (see Fig. 4).

III. Missing Traffic Sign Video Dataset

In this paper, we present a novel dataset, MTSVD. What differentiates the MTSVD from other traffic sign datasets is the inclusion of $2K$ missing sign clips having traffic sign cues (refer Table I). Further, it is the first publicly accessible data for missing objects. The dataset contains $135K$ frames of missing traffic sign intervals spread across $2K$ video clips covering 20 traffic sign classes for identifying, classifying, and detecting missing objects in images and video clips. The data is collected in Asia, covering a wide variety of terrain, weather, lighting conditions, and nighttime scenes. MTSVD is captured with the DDPAI X2S Pro camera, recording the driving at 25 fps, with a resolution of 2560×1440. The MTSVD contains $10K$ unique traffic sign tracks, annotated with multiple attributes like occlusion, tilt, damage, and truncation. It also involves labels identifying non-standard traffic signs, like background-foreground color, shape, category, etc. Sample annotated images are shown in Fig 3, along with distinctive attributes in red. MTSVD is spread across 101

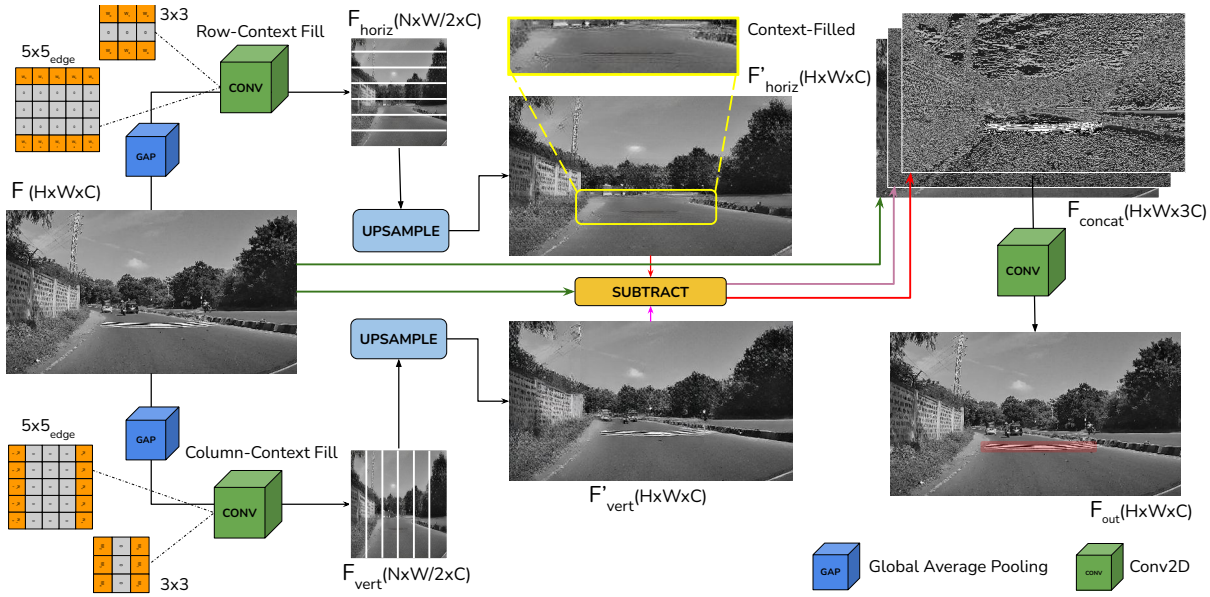

Fig. 4: Illustration of CueCAn. Input feature map, F is downsampled using average pooling. Features are then inpainted using kernel sizes of 3 and 5, with different learnable units (gray: non-learnable, orange: learnable), in a row and column-wise fashion, and upsampled. Features are then subtracted from the original feature map and concatenated with the original feature. Finally, a Conv operation merges the features to obtain the original feature size.

categories of traffic signs.[1] Note that for safety purposes, sometimes the traffic signs appear before the cues in a video sequence (e.g., u-turn). So, we also label the exact intervals when the traffic sign cues are in the camera's field of view.

IV. Methodology

Similar to previous work on non-existent pedestrians [4], we employ VGG-19 [25] encoder followed by the FCN-8 decoder [21] to segment the missing traffic sign regions. We augment VGG-19 layers with novel Cue-driven Contextual Attention (CueCAn) units and train the encoder to classify the presence or absence of traffic signs' cues. We then train the entire network with the enriched encoder features end-to-end for the segmentation task. We now discuss the components of our system in detail.

A. Introducing CueCAn

We model the context space of traffic sign cues as regions with discontinuities. E.g., in the case of speed breakers, or go-slow stripes on the road, this discontinuity is in the form of horizontal ridges. At the same time, the cues for gap-in-median or side-roads (see Fig 2) are complex discontinuities. We model such complexities by a composition of horizontal and vertical discontinuities in feature space. We propose highlighting these aberrations to focus on traffic signs' cues using CueCAn. The CueCAn learns to fill the rows and columns of the image features with their context and finds the difference from the original feature vector. Non-filled regions with cues in the original features have a larger difference than their filled counterparts. The difference helps highlight the cues. Context cues without a linear geometry, like in the case of curves, benefit from the composition of horizontal and vertical filling. The implementation of the CueCAn is illustrated in Fig. 4. First, the input feature map F with shape $[H, W, C]$ is fed to an average pooling layer, condensing the feature to the shape of $[N, W/2, C]$. Here, H, W, and C are the input feature's height, width, and channel dimensions, and N is fixed to 8, similar to Han et al. [24]. The features are then filled with the context in a row-wise fashion with a learnable 3×3 convolution operation with the central row fixed to zero (or depending on the encoder's layer: 5×5 convolution with central three rows fixed to zero), resulting in F_{horiz} of shape $[N, W/2, C]$. A similar process is applied in a column-wise manner to fill the feature columns with context using convolutional filters with central columns fixed to zero, resulting in F_{vert} of shape $[N, W/2, C]$. F_{horiz} and F_{vert} are upsampled (using bilinear interpolation) to the dimensions of original features, i.e. $[H, W, C]$, denoted by F'_{horiz} and F'_{vert}, respectively. Finally, the upsampled features, now containing the context encoding, are individually subtracted from the input feature F, and are further concatenated with F, resulting in F_{concat} of shape $[H, W, 3C]$. The F_{concat} is matched to the input feature $[H, W, C]$ using a Conv+ReLU operation. When filled with their context, regions such as the sky and road do not introduce a significant change. However, cue regions change more with the context filling operation, maximizing the difference between the original and the context-encoded feature, which provides supplementary information to the original feature vector by highlighting the traffic signs' cues.

B. Training Data Creation

To train the encoder for cue classification, we randomly sample 12500 unique frames from MTSVD using the annotated intervals and tracks to create four balanced sub-sets. The sub-sets include i) frames containing traffic signs with cues, ii) frames with traffic sign cues and inpainted traffic signs, using a state-of-the-art inpainting technique [26], iii)

[1]The only other traffic sign dataset to include such dense attributes (6) and categories (313) for each annotation is the MTSD [13] but does not enable missing object-related tasks.

frames without any cue but with traffic signs (from categories having no cues in the same frame generally, e.g., school ahead), and iv) frames without any cue or traffic sign. Thus we balance the dataset to avoid any bias toward the traffic signs or cues. The traffic sign cue is a central part of identifying missing traffic signs, so we only use the inpainted images and corresponding inpainting masks[2] as input-output pairs for the localization task, similar to previous works citeseeingwhatnotthere,missingpedestrians. It is important to note that despite using inpainted images, locating missing traffic signs is more challenging than previous works because i) MTSVD contains 20 complex traffic sign cues, ii) locating traffic signs is challenging even after attending to the cue regions due to the variety in possible sign placements.

C. Model Training

To identify missing signs, the model must attend to the traffic sign cues in the environment. Similar to Han et al. [24], we add CueCAn at the end of the third, fourth, and fifth blocks of the VGG-19 [25] encoder to highlight and classify the cues. The next task is to localize *where* the sign could be placed, using segmentation model with the pre-trained VGG-19 encoder and FCN-8 [21] decoder. For localization, optimal results and GradCAM visualizations are observed when the entire network is fine-tuned end-to-end. We use binary cross-entropy loss for classification and focal loss [27] to handle class-imbalance in the localization task.

V. Experiments and Results

We use a batch size of 32 and train the cue classifier with Adam optimizer with an initial learning rate of $1e^{-4}$. The segmentation model is also trained with Adam optimizer and an initial learning rate of $1e^{-3}$. All models are trained on a single Nvidia RTX-2080Ti GPU for 400 training epochs, with train:val:test split of 80:10:10.

We consider the Vanilla VGG-19 encoder to be the baseline cue classifier. We also experiment with two different versions of the CueCAn, i) by changing the convolutional kernel size and ii) by changing the learnable parameters of the convolution filter. We try convolutional kernel sizes of 3, 5, and 7, with two context-filling approaches for each kernel size. We start with an implementation in which all kernel parameters except the central row/column are learnable. We refer to this configuration as CueCAn$_k$ where k is the kernel size. In the next version, CueCAn$_{ke}$, only the boundary (or edge) rows or columns are learnable. The intuition behind the CueCAn$_{ke}$ is that if only boundary features are used, it leads to better context filling in the feature space. Using only the edge parameters reduces the noise of nearby pixels (or elements in feature space) containing cues due to the receptive field. We verify the intuition mentioned above by the comparative GradCAM [28] visualizations using different configurations, with all other parameters kept constant.

Further, we implement arrangements of CueCAn at the end of all the blocks of the VGG-19 encoder. However, the training loss and accuracy curves indicated the ineffectiveness of this approach. We, thus, implement the final set of experiments with CueCAn in only the $3^{rd}, 4^{th}$ and the 5^{th} blocks of the network, similar to Han et al. [24]. We refer to configurations with convolutional kernels of size k, k', and k'' in blocks 3, 4, and 5 as CueCAn$_{kk'k''}$. Similarly, CueCAn$_{kekek'}$ means that the configurations are similar to CueCAn$_{kkk'}$, but the kernels in blocks $3-4$ are edge-filling convolutions. We first experiment with CueCAn$_{333}$ as a baseline. Motivated by the increasing receptive field in deeper layers, we experiment with CueCAn$_{553}$, CueCAn$_{753}$ to learn to fill more context in the cue regions. However, we observe that due to the receptive field, the contextual regions around the cues also have cue information, making it complex for the non-edge kernels to fill the cues. Hence, we also experiment with CueCAn$_5e53$ and CueCAn$_5e5e3$. As we will see in the next section, using kernels above 5 degrades both precision and recall, we, therefore, avoided using kernels of dimension 7 in the final configurations.

TABLE III: Traffic Sign Cue Classification Results.

Model	Precision	Recall	F-Score
VGG19 [25]	94.77	87.45	90.96
CueCAn$_{333}$	94.87	93.96	94.41
CueCAn$_{553}$	95.30	92.20	93.72
CueCAn$_{753}$	92.48	90.99	91.73
CueCAn$_{5e53}$	96.05	**94.49**	95.26
CueCAn$_{5e5e3}$	**97.96**	93.22	**95.53**

Classification Results: We present the traffic sign cue classification results in Table III. It can be observed that the baseline VGG19 has precision, recall, and F-score of 94.77, 87.45, and 90.96. Using VGG19 with CueCAn$_3$33 shows significant improvements of over 3.5% in Recall and F-score. This is an impressive result since the model with CueCAn learns to better classify multiple types of traffic sign cues (20 categories as discussed in Sec. III) without any cue-level supervision. The classifier only uses a binary label (i.e., presence or absence of a cue) at the frame level. Increasing the kernels' size of context filling filter from 3 to 5 in CueCAN$_{553}$ further improves the precision but reduces the recall as shown in row 3 of Table III. The next row shows that increasing the kernel size to 7 leads to the degradation of all three scores. However, using edge-kernels instead of kernels with only central row fixed to zero (see Sec. IV), leads to further improvements in precision and F-scores as the last two rows of Table III depict, though the recall of CueCAN$_{5e53}$ is better than CueCAN$_{5e5e3}$. Nevertheless, through qualitative analysis in Fig. 6, it can be observed that CueCAN$_{5e53}$, which is qualitatively also better than VGG-19 and CueCAN$_{333}$, fails to focus on the exact cue regions for *gap-in-the median* (row two). Moreover, the last column of Fig. 6 shows that CueCAN$_{5e5e3}$ attends the correct traffic sign cues with high precision compared to the other variants. Thus, we use CueCAN$_{5e5e3}$ for the segmentation task and refer to it as CueCAN henceforth.

Localization Results: Table IV lists the recall rate for

[2]as the ground truth for missing traffic signs doesn't exist.

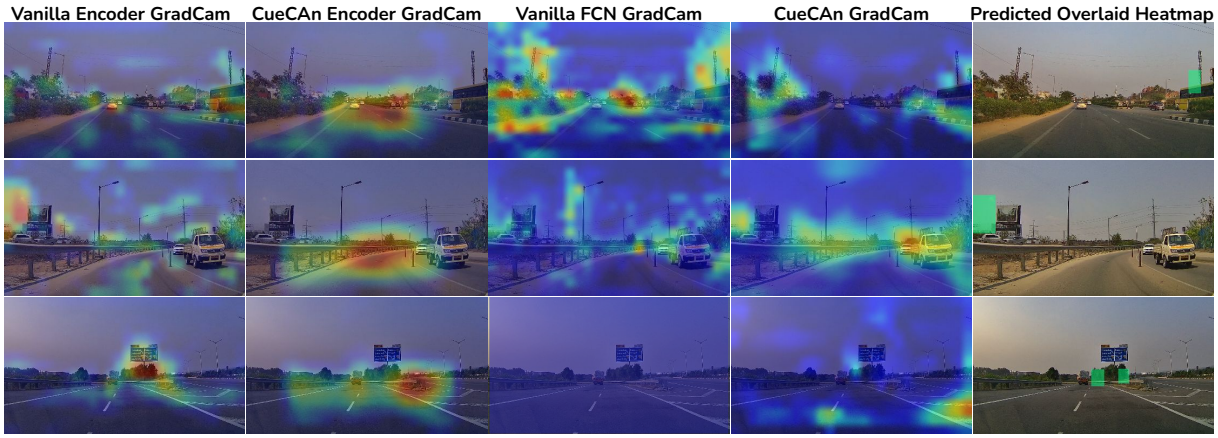

Fig. 5: Illustration for localization of missing traffic signs. **Top**: cue for gap-in-median is attended well by the CueCAn. Overlaid predicted missing sign segmentation map (green mask) is placed rightfully at the gap's neck. **Mid**: cue for a right-hand-curve. The CueCAn kernels and feature compositions help attend to curves. The predicted segmentation map is rightly placed on the left side. **Bottom**: cue for gap-in-median, CueCAn places two signs, one very close to the junction and the other a bit farther away. Previous works fail to localize any sign for all three samples.

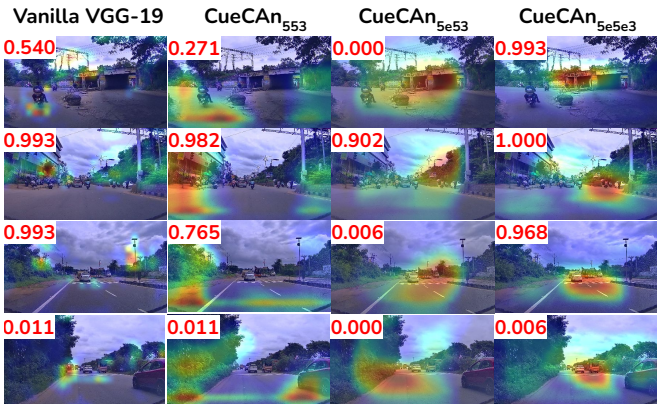

Fig. 6: **Top-to-Bottom**: cues for *height-limit*, *gap-in-median*, *pedestrian-crossing*, and a failure case. The baseline (vanilla VGG19 encoder) fails to attend to context cues. CueCAn_{353} though do attend to cues much better than the baseline, is observed to have low precision towards the important road regions. Using the edge-kernels in the network leads to better results.

TABLE IV: Localization Results

	SFC [5]	FCN [21]	FCN-D [4]	CueCAn	FCN-P	CueCAn-P
Rec.	34.72	47.75	51.23	64.95	71.41	**86.82**

different methods, as proposed by Chien et al. [4] to localize non-existing pedestrians. SFC achieves the lowest recall of 34.72 since it is designed to localize missing objects with cues around them. Such phenomena occur only for *obstacle-delineator* and *bus-bay* traffic signs in our data among 20 categories of missing signs. FCN and FCN-D proposed by Chien et al. [4] achieve better recall rates of 47.75 and 51.23. FCN, with the proposed CueCAN units, achieves the highest recall of 64.95 compared to the previous approaches. We also post-process the predictions from the two localization models (FCN and CueCAN) by taking the tight rectangular region around the predicted blobs. As the last two columns of Table IV depict, the two models with post-processing, i.e., FCN-P and CueCAN-P, significantly improve the recall rates (note FCN-P is better than FCN-D), helping us achieve the recall of 86.82. The result is impressive since GradCAM visualizations of the point at the center of predicted green rectangles in Fig. 6 show that CueCAN attends to the correct traffic sign cues while simultaneously localizing the corresponding signs. We observe that the previous works fail to predict any mask for all Fig. 6 samples (see yellow mask in Fig. 1 (right)). The failure case for our approach is shown in the third row of Fig. 6, which predicts two traffic signs for a single cue, perhaps due to its distant location.

TABLE V: Results for Missing Traffic Sign Video Recognition

Task	Precision	Recall	F-Score
Region classification	59	60	59.49
Video Recognition	37.50	50	42.85

Results on Missing Traffic Sign Videos: We use the CueCAn-P model for video recognition on $2K$ missing sign intervals in MTSVD. We also add $2K$ video clips without any missing sign cue to fairly test our model. We empirically observed that traffic sign predictions near the frame's central column are confusing cases as they represent cues which are far from the camera. Therefore, we use the predictions' centre, height, width, distance from the image center and aspect ratio, and train a Random Forest to classify predicted regions into missing or non-missing (80:10:10 split). Finally, we use majority voting from all CueCAn-P's predictions in an interval for video recognition. Table V results show that missing traffic sign video recognition remains challenging.

VI. Conclusion and Future Work

We presented the Missing Traffic Signs Video Dataset (MTSVD), the first publicly accessible, most complex dataset for missing objects. MTSVD also contains $10K$ traffic sign tracks and $2K$ clips ($135K$ frames) of missing traffic signs. Further, we propose a solution to identify missing signs by training a CueCAn-based VGG-19 cue classification encoder coupled with the FCN-8 decoder to locate missing traffic signs. CueCAn fills the rows and columns of features with their context and subtracts the filled and the original features to highlight the traffic-sign cues. CueCAn significantly improves the results, qualitatively and quantitatively. In the future, we would like to explore MTSVD for missing object detection (with multiple categories) and tracking to investigate further the problems related to missing objects.

Manuscript 2450 submitted to 2023 IEEE International Conference on Robotics and Automation (ICRA). Received September 16, 2022.

REFERENCES

[1] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three Decades of Driver Assistance Systems: Review and Future Perspectives," *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 4, pp. 6–22, 2014. 1

[2] D. Petrović, R. Mijailović, and D. Pesić, "Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers," *Transportation research procedia*, vol. 45, pp. 161–168, 2020. 1

[3] A. F Magnussen, N. Le, L. Hu, and W. Eric Wong, "A survey of the inadequacies in traffic sign recognition systems for autonomous vehicles," *International Journal of Performability Engineering*, vol. 16, no. 10, 2020. 1

[4] J.-T. Chien, C.-J. Chou, D.-J. Chen, and H.-T. Chen, "Detecting Nonexistent Pedestrians," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 182–189. 2, 3, 4, 6

[5] J. Sun and D. 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*, 2017, pp. 5716–5724. 2, 6

[6] K. Hara, J. Sun, R. Moore, D. Jacobs, and J. Froehlich, "Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 189–204. 2

[7] E. Chian, W. Fang, Y. M. Goh, and J. Tian, "Computer vision approaches for detecting missing barricades," *Automation in Construction*, vol. 131, p. 103862, 2021. 2

[8] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 520–527, 2007. 2, 3

[9] M. Bar, "Visual Objects in Context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004. 2, 3

[10] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The Role of Context for Object Detection and Semantic Segmentation in the Wild," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 891–898. 2, 3

[11] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1285–1292. 2

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. 2

[13] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, "The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale," in *European Conference on Computer Vision*. Springer, 2020, pp. 68–84. 2, 3, 4

[14] A. Youssef, D. Albani, D. Nardi, and D. D. Bloisi, "Fast Traffic Sign Recognition Using Color Segmentation and Deep Convolutional Networks," in *International conference on advanced concepts for intelligent vision systems*. Springer, 2016, pp. 205–216. 2, 3

[15] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-Sign Detection and Classification in the Wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118. 2, 3

[16] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. 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. 2, 3

[17] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool, "Traffic sign recognition—how far are we from the solution?" in *The 2013 international joint conference on Neural networks (IJCNN)*. IEEE, 2013, pp. 1–8. 3

[18] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," *Machine vision and applications*, vol. 25, no. 3, pp. 633–647, 2014. 2

[19] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An Empirical Study of Context in Object Detection," in *2009 IEEE Conference on computer vision and Pattern Recognition*. IEEE, 2009, pp. 1271–1278. 3

[20] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8. 3

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 3, 4, 5, 6

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. 3

[23] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3089–3098. 3

[24] X. Han, C. Nguyen, S. You, and J. Lu, "Single Image Water Hazard Detection using FCN with Reflection Attention Units," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 105–120. 3, 4, 5

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4, 5

[26] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 2149–2159. 4

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988. 5

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626. 5