Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SafeSignCue: Evaluating abstaining traffic-sign recognition for assistive cueing

Oded Asiedu

odedasiedu1@gmail.com

Kwame Nkrumah University of Science and Technology (KNUST)

Rose-Mary Gyeming

Kwame Nkrumah University of Science and Technology (KNUST)

Kate Takey

Kwame Nkrumah University of Science and Technology (KNUST)

Emmanuel Ahene

Kwame Nkrumah University of Science and Technology (KNUST)

Justice Agyemang

Kwame Nkrumah University of Science and Technology (KNUST)

Jerry John Kponyo

Kwame Nkrumah University of Science and Technology (KNUST)

Dickson Marfo

Kwame Nkrumah University of Science and Technology (KNUST)

Research Article

Keywords: Assistive navigation, Visual impairment, Selective classification, Uncertainty calibration, Traffic-sign recognition, Evaluation methodology

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11448038/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluating abstaining traffic-sign recognition for assistive cueing

Oded Asiedu[1*], Rose-Mary Gyeming[1], Kate Takyi[1], Emmanuel Ahene[1], Justice Agyemang[2], Jerry John Kponyo[2], Dickson Marfo[2]

[1*]Department of Computer Science, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana.
[2]Department of Telecommunication Engineering , Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana.

*Corresponding author(s). E-mail(s): odedasiedu1@gmail.com;
Contributing authors: rmo.mensah@knust.edu.gh; katetakyi@gmail.com; aheneemmanuel@knust.edu.gh; justiceowusuagyemang@gmail.com; jjkponyo@gmail.com; dmfosu1@st.knust.edu.gh;

Abstract

Purpose A camera-based aid that speaks traffic-sign information to blind and visually impaired (BVI) pedestrians must decide what a sign is and when to stay silent. This study examines how protected error categories and calibration evidence shape reported risk and model choice for such an abstaining recogniser.

Methods Four mobile-scale backbones were trained on the German Traffic Sign Recognition Benchmark with three seeds each. Confidence and entropy thresholds were selected on calibration data to keep wrong Stop, Yield and Crossing cues at or below 0.1% of frames, and repeated for {Stop, Crossing}. Seven image conditions were tested.

Results Clean accuracy spanned under one point (97.28–98.00%), yet calibration recall of protected signs, driven by occluded images, ranged from 82.67% to 88.78%, lowest for the most accurate model. Adding Yield changed the selected thresholds in 11 of 12 runs; reported blur wrong-cue rates were 0.037–0.145% without Yield and 0.322–0.929% with it. MobileNetV3-large was selected under both. Its test wrong-cue rate exceeded 0.1% under blur, overexposure and rotation, and marginally under occlusion. No operating point selected for the three-category set demonstrated the target with 95% confidence: with 3 150 calibration images this requires zero wrong cues, and each had three to six.

Conclusion Clean accuracy did not identify the best recogniser under occlusion, wrong-cue rates are interpretable only with their category set, and the 0.1% target was not demonstrated. No safety requirement is shown to be met; category sets and targets need validation with orientation-and-mobility specialists and BVI users.

Keywords: Assistive navigation, Visual impairment, Selective classification, Uncertainty calibration, Traffic-sign recognition, Evaluation methodology

1 Introduction

Independent travel is one of the most demanding everyday tasks for people who are blind or visually impaired, and mobility-related accidents are common in this population [25]. Camera-based aids that describe the surroundings through speech or haptics are therefore an active research area [20, 38]. Such aids face a constraint that general-purpose recognisers do not: a wrong spoken statement about the traffic environment can be worse than no statement, because the listener often cannot check it. Blind users report that errors of AI-based visual assistance are hard to verify without sight [2], and their acceptance of navigation-aid errors depends on the context in which an error occurs [1]. An assistive recogniser should therefore answer only when its evidence is reliable and otherwise ask for a better view.

Two tasks are easily conflated. The first is recognising signs that are potentially informative to a pedestrian, such as a stop or yield sign at an approaching junction. The second is perceiving information that actually supports a crossing decision, such as the state of a pedestrian signal or the presence and geometry of a crosswalk [37, 43]. This study addresses only the first task, and only for a cropped sign image: it performs no sign detection in a full camera frame, no tracking, no distance or orientation estimation and no judgement of whether a sign applies to the user's path. The recogniser studied here, SafeSignCue, is therefore a candidate component of a future assistive system rather than a navigation aid. It is evaluated on driver-facing German signs without user involvement, and none of the results should be read as evidence about pedestrian safety or about effectiveness for BVI users.

Abstention after training is cheap to add to any classifier, but it moves the burden onto evaluation. The quantity to be controlled has to be defined (which errors count), a threshold has to be chosen on data the test set does not see, and a claim that the error rate is "at most" some value has to be supported by a finite sample. Each choice can change what is reported. We ask three questions. (i) Does clean accuracy predict how many pedestrian-relevant signs a calibrated, abstaining recogniser still announces correctly? (ii) How strongly do the selected thresholds, the reported error rates and the preferred backbone depend on which categories the error constraint protects? (iii) Can a small wrong-cue operating target be demonstrated with a calibration sample of the size available here, and does it hold under controlled image corruptions?

The contributions are methodological and are demonstrated on four mobile-scale backbones. (1) Clean accuracy did not predict calibrated recall of pedestrian-relevant signs, which in this setting is driven by how each model's confidence behaves under occlusion. (2) The definition of the protected category set changed the selected thresholds in almost every run and the reported blur wrong-cue rate by a factor of 4 to 25, so a reported rate is meaningful only together with its category set; for the two definitions compared, the selected backbone did not change. (3) The evidential limits of a small operating target are quantified: with 3 150 calibration images, a 0.1% target can be demonstrated only by an error-free operating point, and no operating point selected for the three-category set was error-free. (4) Complementary risk measures, namely the share of spoken protected cues that are wrong and the share of protected signs that are missed, are reported alongside the per-frame rate, and they show that missed signs dominate under blur and occlusion. The rankings reported here are descriptive results for four models and three seeds each, not tests of statistical superiority.

2 Related work

2.1 Assistive perception for BVI pedestrians

Reviews of navigation support for BVI people cover sensing, localisation and feedback modalities [20], and computer-vision scene understanding for these users has grown rapidly [38]. A systematic mapping of accessible wayfinding systems found that about a third of the reviewed studies involved users neither in design nor in evaluation [31], a gap that the present study shares and states explicitly (Section 5.1). BVI-specific data collection has focused on sidewalk obstacles [36] and crosswalk scenes [37], pedestrian traffic lights and zebra crossings have been recognised with compact networks [43], and beacon-based systems address indoor wayfinding [6]. How spoken guidance should be worded for blind pedestrians is

itself a design question [10]; here the cue phrases are deliberately restricted to observations.

2.2 Selective classification and risk control

The trade-off between error and rejection was formalised by Chow [7]. Selective classification for deep networks abstains when a confidence score falls below a threshold, and the selection-with-guaranteed-risk procedure of Geifman and El-Yaniv chooses that threshold so that a binomial upper bound on the selective risk meets a target with high probability [11]. Rejectors can also be trained jointly with the classifier [12], or, as here, thresholds can be applied after training to scores such as the maximum softmax probability [17]; selective performance across operating points is commonly summarised by risk–coverage curves [13]. Softmax confidence is often miscalibrated [14] and degrades under distribution shift [16]. Distribution-free risk control selects thresholds so that a finite-sample upper confidence bound on the risk, rather than its point estimate, satisfies the target [3, 5], and Learn-then-Test frames the search over many candidate thresholds as a multiple-testing problem [4]. Table 1 contrasts these approaches with the present study, which does not propose a new risk-control procedure: it uses a point-estimate selection rule, as is common in applied work, and then asks what the calibration evidence does and does not support.

2.3 Traffic-sign recognition

Classification on the German Traffic Sign Recognition Benchmark (GTSRB) [33] is close to saturated: MicronNet reaches 98.9% with 0.51 million parameters [42], and a conditional vision transformer reports 99.87% together with a confidence threshold below which outputs are withheld [28]. Neither reports category-specific wrong-cue rates or behaviour under image corruption, and their resolution, training recipe and augmentation differ from ours, so accuracy figures are not directly comparable. To our knowledge no classification-format traffic-sign dataset collected from the pedestrian's viewpoint exists, so this study necessarily uses GTSRB, whose signs are addressed to drivers.

3 Materials and methods

3.1 Task definition and scope

The input to SafeSignCue is a cropped image of a single traffic sign, as provided by GTSRB. The recogniser predicts one of 43 sign classes, maps the prediction to one of nine cue categories and either speaks a fixed observational phrase for that category (for example, "A stop sign is visible.") or, if the uncertainty gate rejects the prediction, says "Sign detected but unclear, please re-aim." (Fig. 1). Phrases are deliberately observations rather than instructions: turning an observation into advice ("wait", "cross now") would require information about direction, distance and path relevance that the component does not have. For Speed signs the phrase names the predicted fine class (for example, "A speed limit 80 sign is visible."), so a wrong fine class within that category is also spoken.

3.2 Dataset and partitioning

GTSRB contains 43 classes. The torchvision [29] GTSRB class used here downloads the archive `GTSRB-Training_fixed.zip`, which provides 26 640 training images, fewer than the 39 209 of the original benchmark release [33], together with the official 12 630-image test set. Training images come from short video tracks of the same physical sign (30 frames per track), so the training pool was split by (class, track) into training, validation and calibration partitions of 20 340, 3 150 and 3 150 images. Whole tracks were assigned per class until at least 10% of the class's images had been placed in the validation partition and then in the calibration partition; because tracks are indivisible, each of these partitions holds 11.8% of the pool. No physical sign contributes frames to more than one partition. The split was generated once (split seed 6266324) and reused for every run. The validation partition was used only for checkpoint selection during training and the calibration partition only for threshold selection, backbone selection and the temperature-scaling diagnostic; the official test set was evaluated once per run after the thresholds had been fixed.

The protected categories are sparse in the calibration partition. From the recall values in the selection records, each calibration copy contains 150 Stop and Crossing frames and 150 Yield

Table 1 Position of this study relative to closely related work. The entries describe what each cited work does in the respect named by the column

Work	Setting	Errors controlled	Threshold selection	Finite-sample guarantee
Geifman and El-Yaniv [11]	Image classification	Selective (misclassification) risk	Binomial-bound search over confidence thresholds	Yes, for the selected threshold
Bates et al. [5]; Angelopoulos et al. [3]	General prediction sets	User-defined loss	Calibrated threshold on a held-out set	Yes
Angelopoulos et al. [4]	General	User-defined risks	Multiple testing over a grid of parameters	Yes, accounting for the search
Naz et al. [28]	GTSRB traffic signs	None reported	Fixed confidence threshold (0.8)	No
This study	GTSRB crops, spoken cues for BVI pedestrians	Wrong cues within a stated protected category set	Point-estimate rule on calibration data, two category sets	Assessed post hoc at the selected points; not met

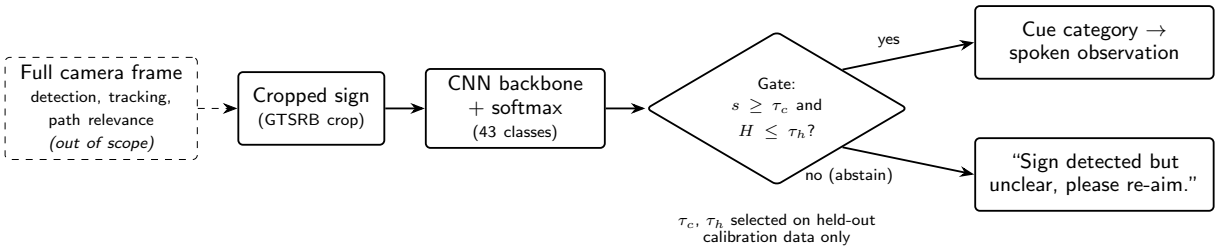


Fig. 1 Scope of SafeSignCue. The study evaluates the solid-outlined steps on cropped signs; detection in the full camera frame, tracking and judgement of path relevance (dashed) are outside its scope. s: maximum softmax probability; H: predictive entropy. The thresholds τ_c and τ_h are selected on held-out calibration data only

frames, that is, about five physical sign tracks of each; the official test set contains 480 Stop and Crossing images and 720 Yield images. The split of the 150 frames between Stop and the two Crossing classes is not recorded in the released artifacts.

3.3 Cue taxonomy and protected category sets

Table 2 maps the 43 classes to nine cue categories. Two protected category sets are compared:

$$C_0 = \{\text{Stop},\ \text{Crossing}\}, \tag{1}$$

$$C_1 = \{\text{Stop},\ \text{Yield},\ \text{Crossing}\}. \tag{2}$$

C_0 is the set that the cue taxonomy labels "high urgency". C_1 adds Yield, which marks a location where crossing traffic has priority. Neither set has been elicited from BVI pedestrians or validated by orientation-and-mobility (O&M) specialists, and three caveats apply. First, Yield's primary addressee is the driver. Second, the two Crossing classes ("pedestrians", "children crossing") are driver-facing warnings, so a correct Crossing cue means only that drivers are warned about pedestrians at this place, not that crossing is safe. Third, Warning, which mixes pedestrian-relevant classes such as "traffic signals" with irrelevant ones such as "slippery road", and Priority are defensible additions that are not evaluated here.

History of the category sets

The project first selected thresholds for C_0 and evaluated them on the test set. According to the project records (the docstring of the recalibration script in Online Resource 1), C_1 was introduced afterwards, when a diagnostic analysis of C_0 test predictions under blur showed frequent wrong Yield cues that C_0 did not count. The thresholds for C_1 were then selected on the calibration partition alone and evaluated once on the test set, but the decision to study C_1 was informed by test-set results. The comparison of C_0 and C_1 is therefore post hoc, and C_1 results are not independent of the test set at the level of the category definition. Earlier development runs, which used hand-chosen thresholds and are not reported here, preceded both protocols.

Table 2 Mapping of the 43 GTSRB classes to the nine cue categories. Crossing comprises classes 27 (pedestrians) and 28 (children crossing); Priority comprises 11 (right-of-way at next intersection) and 12 (priority road). C_0: Stop and Crossing; C_1: Stop, Yield and Crossing

Cue category	Urgency	GTSRB class indices
Stop	High	14
Crossing	High	27, 28
Yield	Caution	13
Warning	Caution	18–26, 29–31
Prohibition	Caution	9, 10, 15–17
Mandatory	Caution	33–40
Speed	Info	0–5, 7, 8
Priority	Info	11, 12
Info	Info	6, 32, 41, 42

3.4 Backbones and training

MobileNetV3-small and MobileNetV3-large [18], EfficientNet-Lite0 [35] and ResNet18 [15] were instantiated from the timm library [40] as ImageNet-pretrained [32] trunks with global pooling, followed by a new head $D \to 256 \to 43$ (Hardswish, dropout 0.2), where D is the pooled feature width of the trunk (1 024 for MobileNetV3-small, 1 280 for MobileNetV3-large and EfficientNet-Lite0, 512 for ResNet18). Images were resized to 96×96 pixels. Training had two stages: four epochs in which only the head was optimised (learning rate 3×10^{-3}), then 24 epochs of full fine-tuning (learning rate 4×10^{-4}). Both stages used AdamW [24] (weight decay 10^{-4}, batch 128) with linear warm-up over 5% of the steps followed by cosine decay [23], label smoothing 0.1 [34], gradient clipping at norm 1.0 and an exponential moving average (EMA, decay 0.999) [30] of the weights. Augmentation comprised resizing to 104×104 with a random 96×96 crop, TrivialAugment [27] and random erasing [45] (probability 0.25, 2–10% of the image area); MixUp [44] ($\alpha = 0.2$) was applied in the second stage only. Horizontal flips were not used because many signs are orientation-specific. All settings are listed in Table A1.

Checkpoints were selected on EMA validation accuracy, and the EMA weights were chosen as the released model before training. Within each stage, the EMA weights of the epoch with the highest EMA validation accuracy were retained; of the two stages, the one whose retained weights scored higher was selected. Four implementation details matter for interpretation. (i) In the first stage the trunk parameters were not updated, but the network was in training mode, so batch-normalisation statistics of the trunk were updated. (ii) With decay 0.999 over the 636 steps of the first stage, the averaged head still carries about half of its random initialisation, which explains the low first-stage EMA accuracy (Table A2); the second stage starts from these averaged weights, and its EMA is reset. (iii) The best EMA validation accuracy was reached in epoch 23 or 24 of the 24-epoch second stage in every run, so validation accuracy was still rising and training may not have fully converged. (iv) The best non-averaged validation accuracy of the second stage exceeded the best EMA accuracy in 10 of 12 runs, by 0.06–0.86 points. Each backbone was trained with three seeds (6266324–6266326), which change only initialisation, batch order and augmentation sampling. Appendix A documents a training fault found and corrected during development.

3.5 Uncertainty gate and threshold selection

For softmax probabilities $\mathbf{p} = (p_1, \ldots, p_{43})$ the gate uses the maximum-softmax confidence $s = \max_k p_k$ and the predictive entropy $H = -\sum_k p_k \log p_k$ (in nats; its maximum is $\ln 43 \approx 3.76$). A cue is spoken when $s \geq \tau_c$ and $H \leq \tau_h$; otherwise the system abstains. Temperature scaling (Section 4.5) is a diagnostic only and is not part of the gate.

For each protected set, thresholds were selected separately for every run on the calibration partition, evaluated twice, clean and with the centre-occlusion corruption of Section 3.6 (6 300 frames from 3 150 images), over the grid $\tau_c \in \{0.10, 0.15, \ldots, 0.95\}$ and $\tau_h \in \{0.50, 0.75, \ldots, 3.50\}$ nats. The rule maximises protected recall, defined as the fraction of frames whose true category is protected that are accepted and assigned the correct category, averaged over the clean and occluded copies, subject to a pooled wrong-cue rate of at most 0.1%. A wrong cue is an accepted frame whose predicted category is protected but differs from the true category. This rate is computed over all evaluated frames, of which about one in ten has a protected true category under C_1; it is not a rate per spoken protected cue, which is reported separately (Section 3.7). If no grid point satisfied the constraint, the one with

the lowest wrong-cue rate would be taken; this fallback was never needed. Ties in recall are broken towards the lowest τ_c and then the lowest τ_h.

Three properties of this rule matter for interpretation. The constraint is discrete: with 6 300 pooled frames, 0.1% allows at most six wrong cues, so the selected threshold can hinge on a single error. The constraint applies to the pooled mixture of clean and occluded frames and implies no bound for either condition separately. And the grid has a floor: its lowest confidence value was selected in all 12 runs under C_0 and in 4 of 12 under C_1. Because ties are broken towards low thresholds, a flat recall plateau at the bottom of the grid produces the same observation, so these selections do not locate an unconstrained optimum.

Operating target

The 0.1% figure is an operating point fixed before evaluation. It is not derived from a risk model, a regulatory requirement or user research, and meeting it would not establish safety; it serves to compare backbones under a common, strict constraint. Because the rule accepts a threshold whose observed calibration rate is at most 0.1%, it ignores sampling error. Section 4.4 therefore assesses, for the selected operating point of each run only, whether a one-sided 95% Clopper–Pearson [8] upper bound on the underlying wrong-cue probability is at most 0.1%. This assessment is a post hoc diagnostic: it does not account for the search over 234 threshold pairs, which a risk-control procedure such as selection with guaranteed risk or Learn-then-Test would have to do [4, 11], and only the selected point of each run was recorded.

Backbone selection and reuse of calibration data

The recommended backbone is the one with the highest mean calibration recall among the backbones that satisfy the constraint in every seed; test-set results played no part in threshold or backbone selection within each protocol. Because the calibration partition is used both to choose thresholds among 234 pairs and to choose among four backbones, calibration recall is a selection criterion and an optimistic estimate of performance; the test-set results are the relevant estimates.

3.6 Controlled image corruptions

Each test image was evaluated clean and under six controlled corruptions applied at the native image resolution, before resizing: Gaussian blur (radius 2 pixels), darkening (brightness $\times 0.35$), overexposure (brightness $\times 1.8$), low contrast (contrast $\times 0.45$), rotation by 12° (nearest-neighbour resampling, exposed corners filled black) and a black patch covering the central third of the image in each dimension (occlusion, one ninth of its area). Each corruption was applied at a single strength, in the spirit of [16] but without its graded severities. Because GTSRB images differ in native size, a fixed blur radius degrades small images more than large ones. Several corruptions overlap with the training augmentation: TrivialAugment samples brightness, contrast and rotation operations (rotations up to 135°), and random erasing replaces up to 10% of the image with the normalised value zero (the ImageNet mean colour), close to the 11.1% black occlusion patch. Gaussian blur has no counterpart in training; the nearest operation, a reduction of sharpness, applies only a 3×3 smoothing filter. These corruptions probe robustness to controlled degradations of well-cropped images and are not a proxy for footage recorded while walking, which adds motion blur, perspective, scale changes, clutter, real glare and detector-generated crops.

3.7 Risk measures and statistical analysis

Three complementary measures accompany the per-frame wrong-cue rate on the test set: the share of spoken protected cues that are wrong (wrong cues divided by all accepted frames whose predicted category is protected), the share of protected signs that are missed (protected frames not accepted with the correct category, divided by all protected frames) and the share of all frames that are accepted with a wrong fine class, which for Speed signs bounds incorrect spoken content not captured by the category-level measures. All are computed from the released per-run counts.

Per-run metrics are reported as mean ± sample standard deviation over the three seeds. Rankings of backbones are descriptive; no significance tests were performed, and with three seeds per

backbone small differences should not be interpreted. For rates pooled over seeds, the target population is the GTSRB test distribution, treating test images as independent draws. Pooling the 37 890 predictions and applying a two-sided Wilson [41] interval treats them as independent, although the three seeds' predictions on the same image are positively correlated. With $m = 3$ correlated predictions per image the design effect $1 + (m - 1)\rho$ is at most m [19], so we also report a seed-adjusted interval that applies the Wilson formula with the count and the sample size both divided by three; it is exact only if all seeds always agree and is otherwise conservative with respect to seed correlation. It does not account for dependence between images of the same physical sign: test-track identifiers are not part of the official annotation, so this residual clustering cannot be removed, and intervals for rates near 0.1% should be read with that caveat. Wilson intervals are used for two-sided descriptive intervals with large counts; the one-sided Clopper–Pearson bound is used for the target assessment because it is exact and conservative for the small counts involved.

For the calibration evidence, the 6 300 frames are two views of 3 150 images, and frames of the same image are not independent. In addition to a frame-level bound, we therefore report an image-level bound that applies the observed wrong-cue count k to the 3 150 images. This is conservative: the number of images with at least one wrong cue is at most k, the per-frame rate cannot exceed the probability that an image yields a wrong cue in at least one view, and the Clopper–Pearson bound increases with the count. It treats images as independent, which tracks of about 30 frames violate; a track-level analysis would widen the bounds further.

4 Results

4.1 Clean accuracy did not predict calibrated recall

All four backbones reached 97.28–98.00% clean test accuracy despite a six-fold range in parameter count (Table 3), and every backbone met the calibration constraint in every seed. Calibration recall of C_1 categories nevertheless ranged from 82.67% for ResNet18, the most accurate model, to 88.78% for MobileNetV3-large. This difference arises under occlusion: recall on the clean calibration copy was 98.00–100% in every run, whereas recall on the occluded copy was 64.33–81.00% (Table A3), and on the clean test set C_1 recall was high for all four backbones (97.28–98.81%). Under occlusion on the test set, ResNet18 abstained far more often than the other models (26.62%) and still had the lowest protected recall (72.00%), whereas MobileNetV3-large combined the lowest occlusion abstention (4.11%) with the highest recall (80.42%). The rankings by clean accuracy, calibration recall and occlusion recall therefore differ (Fig. 2). With three seeds per backbone and overlapping standard deviations, for example 84.11±2.31% and 86.50±1.32% calibration recall for MobileNetV3-small and EfficientNet-Lite0, these rankings are descriptive.

4.2 The protected category set changed thresholds and reported risk

The same twelve trained models were calibrated and tested under C_0 and under C_1 (Table 4). Adding Yield to the protected set changed the selected threshold pair in 11 of 12 runs and raised the selected confidence threshold in 8 of 12; under C_0 every run selected the lowest confidence value of the grid. The reported risk changed far more than the thresholds. Under blur, the C_0 protocol reported wrong-cue rates of 0.037–0.145%, three of four backbones below the 0.1% target, whereas the C_1 protocol reported 0.322–0.929%, all above it. Because thresholds and counting differ together, and because adding a category to the counted errors can only add errors at fixed thresholds, these figures are not a controlled estimate of the contribution of Yield errors; a re-scoring of identical predictions under both definitions requires per-sample outputs, which are not part of the released artifacts but can be regenerated from the released checkpoints with the inference script in Online Resource 1. What the comparison shows is that the same models, evaluated on the same frames, yield headline error rates that differ several-fold depending on the protocol's category set, so a reported wrong-cue rate is meaningful only together with that set.

The selected backbone did not change between the two definitions. The selection rule chose MobileNetV3-large under both, and

Table 3 Backbone comparison under the C_1 objective (mean ± SD over three seeds; bold: highest mean among the accuracy and recall columns). Calibration metrics are computed on the calibration partition (clean and occluded copies, 6 300 frames per run); abstention and recall on the full 12 630-image test set. Clean accuracy is single-view top-1 accuracy over all test images, irrespective of the gate. Calibration recall is the selection criterion and therefore an optimistic estimate of performance; test-set columns are not used for any selection

		Clean test	Calibration (C_1)		Test, occlusion (C_1)	
Backbone	Params (M)	accuracy (%)	Recall (%)	Wrong cues (%)	Abstention (%)	Recall (%)
MobileNetV3-small	1.79	97.28±0.07	84.11±2.31	0.090±0.009	4.74±6.23	75.00±3.03
MobileNetV3-large	4.54	97.72±0.18	**88.78±1.42**	0.069±0.024	4.11±5.04	**80.42±0.22**
EfficientNet-Lite0	3.71	97.73±0.19	86.50±1.32	0.069±0.009	7.66±9.17	79.28±1.19
ResNet18	11.32	**98.00±0.15**	82.67±0.73	0.079±0.016	26.62±11.01	72.00±1.28

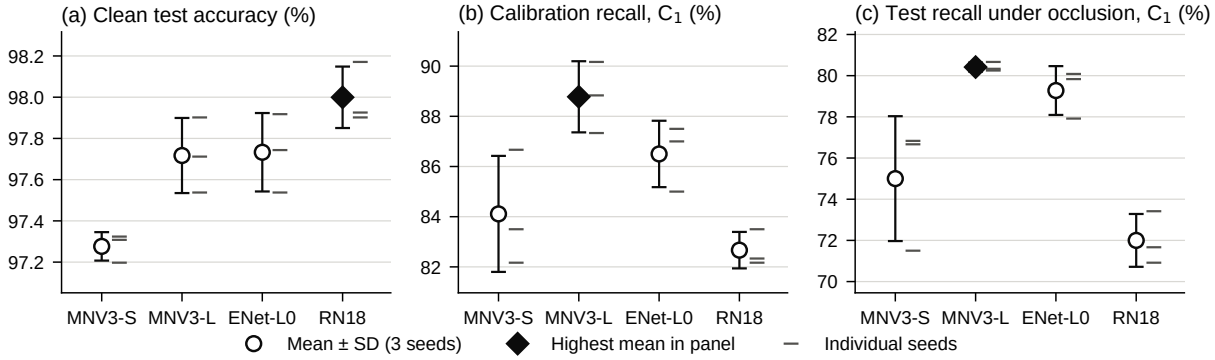


Fig. 2 The four backbones ranked by (a) clean test accuracy, (b) calibration recall of C_1 categories and (c) test recall of C_1 categories under occlusion. Open circles: mean; whiskers: ± SD over three seeds; short horizontal ticks: individual seeds; filled diamond: highest mean in the panel. MNV3-S/L: MobileNetV3-small/large; ENet-L0: EfficientNet-Lite0; RN18: ResNet18. Calibration recall is a selection criterion (Section 3.5)

the ranking of mean calibration recall was identical (MobileNetV3-large, EfficientNet-Lite0, MobileNetV3-small, ResNet18). Whether this holds for broader plausible sets, such as C_1 with Warning or Priority added, was not evaluated, and the comparison is post hoc (Section 3.3).

4.3 Behaviour under controlled corruptions

Table 5 and Fig. 3 give the test-set behaviour of MobileNetV3-large, the selected backbone. Mean coverage stayed above 95% in every condition, although the spread across seeds is large: in the lowest-coverage seed it fell to 90.2% under occlusion and 93.8% under blur. Class accuracy fell to 67.74% under blur and 50.20% under occlusion, so abstention removed some, but far from all, of the additional errors. The wrong-cue rate was 0.111% on clean images and ranged from 0.077% (dark) to 0.596% (blur, about six times the operating target); the point estimate met the target in one of seven conditions. The seed-adjusted intervals were on average 1.76 times as wide as the naive pooled intervals. Under them, the target was met with confidence (upper limit at most 0.1%) in none of the seven conditions and was compatible with the data in three (clean, dark, low contrast). Under blur, overexposure and rotation the whole interval lay above the target; under occlusion the lower limit (0.1005%) exceeded it by less than 0.001 percentage points, and because track-level dependence is not accounted for, occlusion is best described as borderline. Across all four backbones (Tables 5 and A4), the seed-adjusted interval lay entirely below the target in 1 of 28 backbone–condition combinations (EfficientNet-Lite0, dark) and entirely above it in 16, one of them (MobileNetV3-large, occlusion) by the narrow margin just described.

The complementary measures (Table 6; Table A5 for the other backbones) change the picture of what goes wrong. For MobileNetV3-large, 1.17% of the protected cues it spoke on clean images were wrong and 7.04% under blur. Missed

Table 4 The same twelve trained models under two protocols (mean ± SD over three seeds). C_0: thresholds selected to protect, and wrong cues counted over, {Stop, Crossing}; C_1: the same for {Stop, Yield, Crossing}. Each protocol's recall and wrong-cue rate are computed over its own category set, so values are comparable within a protocol, not between protocols; the selected thresholds of every run are listed in Table A3. Wrong-cue rates are on the 12 630-image test set. The selection rule chose MobileNetV3-large under both protocols

	Calibration recall (%)		Clean wrong-cue rate (%)		Blur wrong-cue rate (%)	
Backbone	C_0	C_1	C_0	C_1	C_0	C_1
MobileNetV3-small	71.33±1.86	84.11±2.31	0.119±0.029	0.164±0.016	0.037±0.037	0.929±0.158
MobileNetV3-large	78.78±2.46	88.78±1.42	0.103±0.057	0.111±0.055	0.145±0.025	0.596±0.067
EfficientNet-Lite0	74.00±3.53	86.50±1.32	0.045±0.012	0.053±0.024	0.066±0.025	0.855±0.508
ResNet18	68.78±1.02	82.67±0.73	0.124±0.012	0.113±0.040	0.079±0.032	0.322±0.171

protected signs were the larger failure under blur (17.08%) and occlusion (19.58%): pooled over seeds, 615 protected frames were missed under blur against 226 wrong protected cues, and 705 under occlusion against 59. Under occlusion, 46.36% of all frames were accepted with a wrong fine class, so the per-frame wrong-cue rate, which counts only errors into protected categories, describes a small part of the incorrect spoken output.

Two features of these results bear on how a target set during calibration transfers to testing. First, blur and overexposure, which had the largest excess, are absent from the calibration pool, which contains only clean and occluded images; but occlusion, which is in the pool, also reached the target's boundary, so a mismatch between calibration and test conditions cannot be the only cause. Second, on clean calibration frames the selected thresholds produced 10 wrong C_1 cues in 37 800 frames pooled over all twelve runs (0.026%), and on occluded calibration frames 48 (0.127%); on the clean test set the corresponding figure was 167 in 151 560 predictions (0.110%). Clean-image error rates were thus about four times higher on the official test set than on the calibration partition. Figure 4 shows representative gate decisions, including a case in which abstention discarded a correct prediction.

4.4 Can the calibration sample demonstrate a 0.1% target?

The selection rule accepts a threshold whose observed wrong-cue rate on the calibration pool is at most 0.1%. Because that pool contains 6 300 frames, the rule decides on a handful of events: the selected C_1 operating points of the twelve runs contain between three and six wrong cues (Table 7). The corresponding one-sided 95% upper bounds range from 0.123% to 0.188% at frame level and from 0.246% to 0.376% at image level, so no selected C_1 operating point demonstrates the target. The consequences of this sensitivity are visible downstream: for MobileNetV3-large the three seeds, which differ only in initialisation, batch order and augmentation sampling, selected $\tau_c = 0.20$, 0.15 and 0.10 and consequently covered 90.2%, 97.9% and 99.7% of the test frames under occlusion.

The size of the calibration sample does not by itself rule out a demonstration. At image level, 3 150 images suffice if the operating point produces no wrong cue at all (upper bound 0.095%), and at least 2 995 independent error-free observations are needed for any 0.1% demonstration of this kind; at frame level, up to two wrong cues would suffice (upper bound 0.0999%). Under C_0, five of the twelve selected operating points had no wrong cue and so meet the target by this criterion at their selected point, but C_0 protects only Stop and Crossing. For C_1, only the selected point of each run was recorded, so whether a stricter grid point would have been error-free, and at what cost in recall, cannot be established from the released artifacts; a procedure that selects among many thresholds would in any case have to account for that search [4]. With about five physical tracks per protected category in each calibration copy, the effective sample is also far smaller than the frame count suggests. The 0.1% target should therefore be read as a design aim that this study did not demonstrate.

Table 5 MobileNetV3-large under controlled corruptions (C_1 thresholds; 12 630 test images × three seeds). Accuracy, abstention, rate and recall: mean ± SD over seeds. Wrong/frames: pooled count of wrong C_1 cues. The naive Wilson interval treats the 37 890 predictions as independent; the seed-adjusted interval divides count and sample size by three, the maximum design effect for three correlated predictions per image. Neither interval accounts for dependence between images of the same physical sign (Section 3.7). †Lower limit 0.1005%, above the target by less than 0.001 percentage points

Condition	Class accuracy (%)	Abstention (%)	Wrong/ frames	Wrong-cue rate (%)	Naive 95% CI (%)	Seed-adjusted 95% CI (%)	Recall C_1 (%)
Clean	97.72±0.18	0.07±0.10	42/37 890	0.111±0.055	[0.082, 0.150]	[0.066, 0.186]	98.28±1.13
Blur	67.74±1.24	2.59±3.09	226/37 890	0.596±0.067	[0.524, 0.679]	[0.476, 0.747]	82.92±0.52
Dark	96.78±0.19	0.14±0.17	29/37 890	0.077±0.050	[0.053, 0.110]	[0.041, 0.142]	97.58±0.79
Overexposure	94.95±0.28	0.23±0.27	108/37 890	0.285±0.035	[0.236, 0.344]	[0.206, 0.394]	98.36±1.12
Low contrast	97.31±0.20	0.08±0.09	45/37 890	0.119±0.029	[0.089, 0.159]	[0.072, 0.196]	97.42±0.71
Rotation	94.07±0.19	0.26±0.35	73/37 890	0.193±0.056	[0.153, 0.242]	[0.130, 0.286]	96.47±0.34
Occlusion	50.20±1.03	4.11±5.04	59/37 890	0.156±0.016	[0.121, 0.201]	[0.100†, 0.241]	80.42±0.22

Table 6 Complementary C_1 risk measures for MobileNetV3-large on the test set (three seeds pooled). Per frame: wrong C_1 cues divided by all test frames (the primary metric). Among spoken: wrong C_1 cues divided by all accepted cues whose predicted category is in C_1. Missed: protected frames not accepted with the correct category, divided by the 3 600 protected frames (1 200 per seed). Accepted misclassified: frames accepted with a wrong fine class, as a fraction of all frames (mean ± SD over seeds); for Speed signs the spoken phrase names the fine class, so this is an upper bound on any incorrect spoken content. Brackets: seed-adjusted 95% Wilson intervals

Condition	Wrong per frame (%)	Wrong/spoken protected cues	Wrong among spoken (%)	Missed protected signs (%)	Accepted misclassified (%)
Clean	0.111	42/3 580	1.17 [0.70, 1.96]	1.72 [1.13, 2.63]	2.23±0.12
Blur	0.596	226/3 211	7.04 [5.65, 8.73]	17.08 [15.06, 19.32]	30.04±1.79
Dark	0.077	29/3 542	0.82 [0.44, 1.52]	2.42 [1.69, 3.45]	3.13±0.10
Overexposure	0.285	108/3 649	2.96 [2.15, 4.07]	1.64 [1.06, 2.53]	4.87±0.15
Low contrast	0.119	45/3 552	1.27 [0.77, 2.08]	2.58 [1.83, 3.64]	2.62±0.12
Rotation	0.193	73/3 546	2.06 [1.39, 3.04]	3.53 [2.62, 4.73]	5.76±0.03
Occlusion	0.156	59/2 954	2.00 [1.29, 3.08]	19.58 [17.44, 21.92]	46.36±5.11

4.5 Gate signals and confidence calibration of the selected backbone

For MobileNetV3-large, the entropy signal added nothing measurable to confidence over the coverage range examined (Table 8). Confidence and negative entropy were almost perfectly rank-correlated on the clean test set (Spearman $\rho \geq 0.997$ in every seed), and when each single-signal gate was set to 50%, 70% or 90% coverage the two gates disagreed on at most 1.77% of frames. At these matched coverages the entropy-only gate never had a lower selective risk than the confidence-only gate, and requiring both conditions never lowered the risk relative to confidence alone. The selected operating points, however, lie at 90–100% coverage, and the selected entropy thresholds (3.00–3.50 nats) were the smallest grid values that preserved the maximum recall, so entropy may contribute at those points; this analysis does not establish that the entropy condition is unnecessary there. It was run only for MobileNetV3-large.

Temperature scaling [14], fitted on the clean calibration images, gave $T = 0.57 \pm 0.01$ for MobileNetV3-large (Table 9). Because $T < 1$, the raw softmax outputs are underconfident, which is consistent with reports that label smoothing and MixUp alter the confidence of trained networks [26, 39]. Scaling reduced the clean-test expected calibration error from 18.06% to 0.45% and the negative log-likelihood from 0.276 to 0.085. The underconfidence is consistent with the low selected confidence thresholds (0.10–0.30): raw scores understate empirical accuracy, so a threshold that looks permissive need not be. Temperature scaling was run only for MobileNetV3-large, is not part of the gate and was not evaluated at the selected operating points.

5 Discussion

The first finding is that clean accuracy did not identify the recogniser that announces the most pedestrian-relevant signs correctly once it is calibrated to abstain. The four backbones were separated by less than one point of clean accuracy,

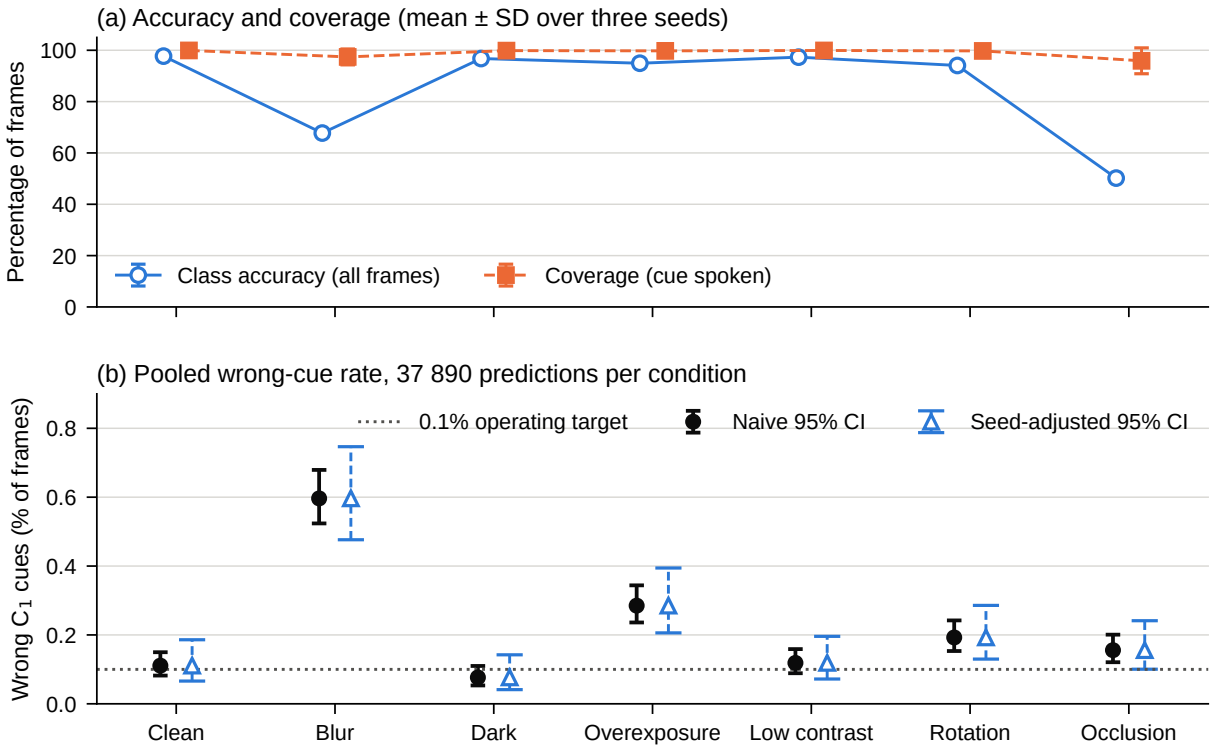


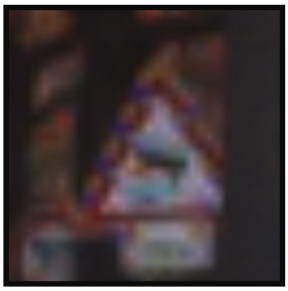
Fig. 3 MobileNetV3-large under controlled corruptions. (a) Class accuracy over all frames (circles, solid line) and coverage, the fraction of frames for which a cue is spoken (squares, dashed line); mean ± SD over three seeds. (b) Pooled wrong C_1 cue rate with naive (filled circles, solid whiskers) and seed-adjusted (open triangles, dashed whiskers with wider caps) 95% Wilson intervals; the dotted line marks the 0.1% operating target. Neither interval accounts for dependence between images of the same physical sign

Table 7 Finite-sample evidence for the 0.1% operating target at the selected thresholds. Counts are wrong cues among the 6 300 pooled calibration frames (3 150 images, each seen clean and occluded) of each seed. Bounds are one-sided 95% Clopper–Pearson upper limits. Frame-level: frames treated as independent. Image-level: the count is applied to the 3 150 images, which is conservative because the number of images with at least one wrong cue cannot exceed the number of wrong cues, and the per-frame rate cannot exceed the per-image rate; it remains valid whether one or both views of an image erred. A run qualifies when its image-level bound is at most 0.1%, which requires zero wrong cues. Only the selected operating point of each run is assessed

	C_1 protocol				C_0 protocol	
	Wrong cues	95% upper bound (%)		Runs	Wrong cues	Runs
Backbone	per seed	Frame-level	Image-level	qualifying	per seed	qualifying
MobileNetV3-small	6, 6, 5	0.167–0.188	0.333–0.376	0 of 3	0, 0, 0	3 of 3
MobileNetV3-large	6, 4, 3	0.123–0.188	0.246–0.376	0 of 3	6, 6, 1	0 of 3
EfficientNet-Lite0	4, 4, 5	0.145–0.167	0.290–0.333	0 of 3	0, 2, 4	1 of 3
ResNet18	5, 6, 4	0.145–0.188	0.290–0.376	0 of 3	0, 2, 5	1 of 3

yet their calibrated recall of protected signs differed by about six points because their confidence behaved differently on occluded images. What decides the usefulness of an abstaining recogniser is how well its confidence separates correct from incorrect predictions on the hard cases, and that has to be measured on held-out data for each model.

The second finding is that the answer to "how often does it say something wrong?" depends on "which errors?". Changing the protected set from {Stop, Crossing} to {Stop, Yield, Crossing} changed the selected thresholds in almost every run and moved the reported blur wrong-cue rate from figures that mostly met the target to figures



(a) Clean

True class: Slippery road

✓ **Accepted; spoken cue:**

“A hazard warning sign is visible.”

Confidence 67.2%, entropy 1.75 nats



(b) Occlusion

True class: No passing for vehicles over 3.5t

✓ **Accepted; spoken cue:**

“A restricted-area sign is visible.”

Confidence 57.7%, entropy 1.74 nats



(c) Blur

True class: Speed limit 80

✗ **Abstained; spoken: “please re-aim”**

Withheld: No vehicles (wrong)

Confidence 16.6%, entropy 3.12 nats



(d) Occlusion

True class: Keep left

✗ **Abstained; spoken: “please re-aim”**

Withheld: Turn left ahead (wrong)

Confidence 15.7%, entropy 3.19 nats



(e) Dark

True class: Roundabout mandatory

✓ **Accepted; spoken cue:**

“A mandatory-direction sign is visible.”

Confidence 70.7%, entropy 1.64 nats



(f) Occlusion

True class: General caution

✗ **Abstained; spoken: “please re-aim”**

Withheld: General caution (correct)

Confidence 15.2%, entropy 3.18 nats

Fig. 4 Six test predictions of MobileNetV3-large, seed 6266324, at that run's C_1 thresholds ($\tau_c = 0.20$, $\tau_h = 3.25$). Solid frame and check mark: prediction accepted and the cue spoken (all three correct); dashed frame and cross: the gate abstained and the system asked the user to re-aim. In (c) and (d) the withheld prediction would have been wrong; in (f) it would have been correct, showing that abstention also costs coverage. Examples were chosen by outcome type: for each requested combination of condition and outcome, the first matching image in a seeded random subset of 1 500 test images was taken. The other two seeds have lower τ_c (Table A3), so the same images may be treated differently by them. Sign images are GTSRB test images [33] after the stated corruption, as seen by the model

that all exceeded it. The complementary measures add a second dimension: under blur and occlusion, missed protected signs and accepted misclassifications outside the protected set were far more frequent than wrong protected cues. For designers and evaluators of assistive systems, the practical implication is that a wrong-cue rate should not be reported, compared across studies or used as a requirement without the category set that defines it and without the corresponding miss rate. Which categories and which kinds of error matter to BVI pedestrians cannot be settled by further analysis of GTSRB; it should be settled with O&M specialists and BVI users, for example through a structured expert elicitation of the pedestrian relevance of each of the 43 classes, followed by evaluation under the elicited set.

The third finding concerns the evidence behind the operating target. The 0.1% point estimate was met in only one of seven test conditions for the selected backbone, and no selected C_1 operating point demonstrated the target with

Table 8 Confidence and entropy as gate signals for MobileNetV3-large (clean test set, mean ± SD over three seeds). Each single-signal gate is set to the target coverage; the selected operating points lie at higher coverage (90–100%), which this analysis does not cover; risk is the misclassification rate among accepted frames. Disagreement: fraction of frames accepted by one single-signal gate but not the other (mean, maximum in parentheses). Combined: both conditions required. Spearman ρ between confidence and negative entropy: 0.9974–0.9977

Target coverage	Disagreement (%)	Selective risk (%)		
		Confidence	Entropy	Combined
50%	1.69 (1.77)	0.095±0.016	0.116±0.009	0.097±0.016
70%	1.46 (1.57)	0.132±0.046	0.158±0.041	0.133±0.046
90%	0.99 (1.22)	0.358±0.094	0.472±0.031	0.360±0.094

Table 9 Temperature scaling of MobileNetV3-large (T fitted by minimising the negative log-likelihood on the 3 150 clean calibration images; metrics on the clean test set; mean ± SD over three seeds). $T < 1$ indicates underconfident raw outputs. ECE: expected calibration error with ten equal-width confidence bins; NLL: negative log-likelihood

Clean test set	Before scaling	After scaling
Temperature T	1 (raw)	0.57±0.01
ECE (%)	18.06±1.54	0.45±0.20
NLL	0.276±0.023	0.085±0.003
Brier score	0.0730±0.0080	0.0355±0.0018

95% confidence. With 3 150 calibration images a demonstration is possible only at an operating point that produces no wrong cue, and the protected categories in each calibration copy come from about ten physical sign tracks, which makes even an error-free result weak evidence. Three consequences follow for studies of this kind. A threshold chosen because its observed error count clears a small target is chosen partly by sampling noise, so a selection procedure based on an upper confidence bound that accounts for the threshold search [4, 11] should be preferred, with a calibration sample large and diverse enough to make the criterion attainable. A target defined on a pooled mixture of conditions does not carry over to each condition. And a target calibrated on one data collection need not hold on another: clean-image error rates were four times higher on the official test set than on the calibration partition.

For MobileNetV3-large, the raw outputs were underconfident, so thresholds that look permissive are not, and entropy added nothing to confidence at the coverage levels examined. Scores that change the ranking of frames, such as deep ensembles of the seeds [21], Monte-Carlo dropout [9] or energy scores [22], are the ones that could improve selective performance.

5.1 Limitations

No BVI participants or O&M specialists were involved, so the protected categories, the cue wording and the operating target are not user-validated; this is the main limitation for an assistive-technology readership, and no claim about pedestrian behaviour, safety or practical effectiveness is made. GTSRB contains German, driver-facing signs; regions with different sign designs would need new data and a revised taxonomy. The recogniser sees ideal crops, whereas a deployed system must first detect the sign, judge its relevance and cope with motion, distance and clutter, and detector-generated crops would be harder than GTSRB crops. The corruptions are controlled, applied one at a time at a single strength, partly overlap with the training augmentation, and "overexposure" is a global brightness change rather than optical glare. Thresholds were calibrated on clean and occluded images only, from about five physical tracks per protected category, and the grid has a floor at $\tau_c = 0.10$ that was selected in many runs. The C_1 category set was adopted after C_0 test results had been examined, only two category sets were compared, and the comparison is not a controlled re-scoring of identical predictions. The seed-adjusted intervals do not account for dependence between images of the same physical sign. The gate-signal and temperature analyses were run for one backbone and not at the selected operating points. Training may not have fully converged, and three seeds per backbone allow only descriptive comparisons. Inference cost was not measured on a mobile device; desktop-GPU timings recorded during the project are not reported because they did not follow model size and are not indicative of on-device cost.

6 Conclusion

This study examined how an abstaining traffic-sign recogniser for assistive cueing should be evaluated, using four backbones as experimental cases. Three conclusions follow, each bounded by the protocol used here. Clean classification accuracy did not identify the model with the highest calibrated recall of pedestrian-relevant signs, which depended on behaviour under occlusion. The definition of the protected category set changed the selected thresholds and the reported wrong-cue rate several-fold, so any reported risk figure must state its category set and should be accompanied by the rate of missed signs; for the two sets compared, the selected backbone, MobileNetV3-large, did not change. And a 0.1% wrong-cue target was neither met under most controlled corruptions nor demonstrated at any operating point selected for the three-category set; with a calibration sample of this size only an error-free operating point could demonstrate it. The recogniser is not shown to meet any safety requirement. The results support further evaluation of calibration-based abstention as a candidate design strategy for assistive cueing, and the category set and the operating target now require validation with O&M specialists and BVI users.

Supplementary information. Online Resource 1 (ESM_1.zip) contains the released per-run artifacts (threshold-selection records, test reports, training histories, temperature-scaling and gate-redundancy measurements for all twelve runs), the training, calibration and evaluation code with the exact commands, a dependency list, the scripts that regenerate every data table and Figs. 2–4 and check every number in the text against the artifacts, and a package manifest with checksums. Online Resource 2 (ESM_2.zip) contains the twelve trained checkpoints with their checksums.

Statements and Declarations

- **Funding** No funding was received for conducting this study.
- **Competing interests** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate** Not applicable. The study uses the public GTSRB dataset and involved no human or animal participants.
- **Consent for publication** Not applicable.
- **Data availability** GTSRB is publicly available at https://benchmark.ini.rub.de/ and was accessed through torchvision. The per-run artifacts from which every data table and figure of this article is computed are provided in Online Resource 1. Per-sample model outputs can be regenerated from the checkpoints in Online Resource 2 with the inference script in Online Resource 1.
- **Materials availability** Not applicable.
- **Code availability** The training, calibration and evaluation code, the exact commands used, a dependency list, an inference script that regenerates per-sample model outputs, and the table, figure and verification scripts are provided in Online Resource 1. The twelve trained checkpoints are provided in Online Resource 2 together with SHA-256 checksums. The code is released under the MIT licence.
- **Author contribution** O.A.: conceptualization, methodology, software, investigation, formal analysis, training audit, writing – original draft. R.-M.G.: [roles]. K.T.: [roles]. E.A.: [roles]. J.A.: [roles]. J.J.K.: [roles]. D.M.: [roles]. All authors read and approved the final manuscript.

Appendix A Supplementary tables and training audit

Table A1 lists the training and evaluation settings and Table A2 the training audit. Table A3 gives the threshold-selection record of every run, Table A4 the corruption results of the three backbones not shown in Table 5, and Table A5 their complementary risk measures.

Training audit

During development, an early run of MobileNetV3-small reached only 20.9% test accuracy. The cause was a software fault: the training script always kept the second-stage checkpoint, and early stopping with a patience of five epochs could end the second stage while its freshly reset

Table A1 Training and evaluation settings, identical for every backbone and seed

Setting	Value
Trunks	timm, ImageNet-pretrained, global pooling
Head	$D \to 256 \to 43$, Hardswish, dropout 0.2
Input	96×96, ImageNet normalisation
Optimiser	AdamW, weight decay 10^{-4}, batch 128
Schedule	linear warm-up (5% of steps), then cosine decay
Stage A	4 epochs, head only, LR 3×10^{-3}
Stage B	24 epochs, full fine-tuning, LR 4×10^{-4}, patience 24
Regularisation	label smoothing 0.1; MixUp $\alpha = 0.2$ (stage B); gradient clipping at norm 1.0
Augmentation	resize 104×104, random 96×96 crop; TrivialAugment; random erasing ($p = 0.25$, 2–10% area)
EMA	decay 0.999, reset at stage B; EMA weights used for selection and evaluation
Split	20 340/3 150/3 150 (train/val/calib), track-grouped, split seed 6266324
Training seeds	6266324, 6266325, 6266326
Threshold grid	τ_c: 0.10–0.95 (step 0.05); τ_h: 0.50–3.50 nats (step 0.25)
Evaluation	single view, 96×96 resize
Training hardware	NVIDIA RTX 4000 SFF Ada Generation GPU

EMA weights still lagged the raw weights. The script was corrected to select between stages as described in Section 3.4, the second-stage patience was raised from five epochs to the full 24-epoch budget (the default in the released script remains five; the commands in Online Resource 1 set it explicitly), and all MobileNetV3-small seeds were retrained before any reported result was produced. The fault was detected through a test-set accuracy that was implausible relative to the other seeds; it was corrected as a software fault, and no hyperparameter was chosen on test results. The training histories of the twelve released runs (Table A2) show that the second stage always ran its full 24 epochs; in one run the EMA validation accuracy did not improve for six consecutive epochs, which a patience of five would have cut short. On the selection quantity, the second stage (best EMA validation accuracy 97.46–98.25%) exceeded the first (43.49–59.24%) by at least 38.22 points in every run, so the stage-selection fault cannot have affected any released checkpoint.

References

[1] Abdolrahmani A, Easley W, Williams M, et al (2017) Embracing errors: examining how context of use impacts blind individuals' acceptance of navigation aid errors. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, New York, pp 4158–4169, https://doi.org/10.1145/3025453.3025528

[2] Alharbi R, Lor P, Herskovitz J, et al (2024) Misfitting with AI: how blind people verify and contest AI errors. In: Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, New York, pp 1–17, https://doi.org/10.1145/3663548.3675659

[3] Angelopoulos AN, Bates S, Fisch A, et al (2024) Conformal risk control. In: International Conference on Learning Representations

[4] Angelopoulos AN, Bates S, Candès EJ, et al (2025) Learn then test: calibrating predictive algorithms to achieve risk control. The Annals of Applied Statistics 19(2):1641–1662. https://doi.org/10.1214/24-AOAS1998

[5] Bates S, Angelopoulos AN, Lei L, et al (2021) Distribution-free, risk-controlling prediction sets. Journal of the ACM 68(6):1–34. https://doi.org/10.1145/3478535

[6] Cheraghi SA, Namboodiri V, Walker L (2017) GuideBeacon: beacon-based indoor wayfinding for the blind, visually impaired, and disoriented. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom), pp 121–130, https://doi.org/10.1109/PERCOM.2017.7917858

[7] Chow CK (1970) On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1):41–46. https://doi.org/10.1109/TIT.1970.1054406

[8] Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the

Table A2 Training audit of the twelve released runs, read from the training histories (validation partition, 3 150 images). Best raw: highest non-EMA validation accuracy within a stage; best EMA: highest EMA validation accuracy, the quantity used for checkpoint and stage selection; released: validation accuracy of the released checkpoint, re-evaluated after training. The stage-A EMA values are low because, with decay 0.999 over 636 steps, the averaged head still carries about half of its random initialisation

Backbone	Seed	Epochs A/B	Best raw, stage A (%)	Best raw, stage B (%)	Best EMA, stage A (%)	Best EMA, stage B (%)	Released (%)
MobileNetV3-small	6266324	4/24	59.43	98.06	54.83	97.71	97.71
	6266325	4/24	57.17	97.17	53.94	97.49	97.49
	6266326	4/24	58.41	97.40	54.32	97.52	97.52
MobileNetV3-large	6266324	4/24	60.22	98.16	59.24	97.46	97.46
	6266325	4/24	59.71	98.60	57.21	97.75	97.75
	6266326	4/24	59.37	98.44	57.87	98.00	98.00
EfficientNet-Lite0	6266324	4/24	57.05	98.76	55.62	98.22	98.22
	6266325	4/24	57.49	98.63	54.29	98.25	98.25
	6266326	4/24	56.38	97.94	53.46	97.81	97.81
ResNet18	6266324	4/24	49.02	97.87	44.03	97.46	97.46
	6266325	4/24	50.06	97.84	45.84	97.68	97.68
	6266326	4/24	49.08	97.71	43.49	97.65	97.65

Table A3 Threshold selection record of every run, read from the released selection files. C_1 columns: selected thresholds, wrong C_1 cues on the clean/occluded calibration copies (3 150 frames each), C_1 recall on each copy (its mean is the selection objective) and coverage of the occluded copy. Narrow columns: thresholds selected for {Stop, Crossing}

		C_1 objective						Narrow objective	
Backbone	Seed	τ_c	τ_h (nats)	Wrong cues clean/occl.	Recall, clean (%)	Recall, occl. (%)	Coverage, occl. (%)	τ_c	τ_h (nats)
MobileNetV3-small	6266324	0.15	3.25	0/6	99.67	67.33	97.27	0.10	2.75
	6266325	0.25	3.00	5/1	100.00	64.33	86.13	0.10	3.00
	6266326	0.10	3.50	0/5	99.67	73.67	99.14	0.10	2.75
MobileNetV3-large	6266324	0.20	3.25	0/6	98.33	76.33	89.62	0.10	3.00
	6266325	0.15	3.25	0/4	99.00	78.67	97.27	0.10	3.25
	6266326	0.10	3.25	1/2	99.33	81.00	99.52	0.10	3.25
EfficientNet-Lite0	6266324	0.10	3.25	0/4	99.00	71.00	96.06	0.10	3.00
	6266325	0.20	3.25	1/3	99.33	74.67	80.86	0.10	3.50
	6266326	0.10	3.25	2/3	98.33	76.67	96.92	0.10	3.00
ResNet18	6266324	0.25	3.25	0/5	98.67	65.67	74.76	0.10	3.25
	6266325	0.20	3.00	1/5	98.00	69.00	86.03	0.10	3.25
	6266326	0.30	3.00	0/4	99.00	65.67	66.29	0.10	3.25

binomial. Biometrika 26(4):404–413. https://doi.org/10.1093/biomet/26.4.404

[9] Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning, pp 1050–1059

[10] Gaunet F (2006) Verbal guidance rules for a localized wayfinding aid intended for blind-pedestrians in urban areas. Universal Access in the Information Society 4(4):338–353. https://doi.org/10.1007/s10209-003-0086-2

[11] Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30, pp 4878–4887

[12] Geifman Y, El-Yaniv R (2019) SelectiveNet: a deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning, pp 2151–2159

[13] Geifman Y, Uziel G, El-Yaniv R (2019) Bias-reduced uncertainty estimation for deep neural classifiers. In: International Conference on Learning Representations

[14] Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330

Table A4 The three backbones not shown in Table 5 under controlled corruptions (C_1 thresholds; mean ± SD over three seeds; Wrong: pooled count of wrong C_1 cues over 37 890 predictions; interval: seed-adjusted 95% Wilson interval of the pooled rate, Section 3.7)

Backbone	Condition	Class accuracy (%)	Abstention (%)	Wrong	Wrong-cue rate (%)	Seed-adjusted 95% CI (%)	Recall C_1 (%)
MobileNetV3-small	Clean	97.28±0.07	0.12±0.19	62	0.164±0.016	[0.107, 0.251]	98.81±0.05
	Blur	67.47±2.13	2.42±2.99	352	0.929±0.158	[0.776, 1.112]	78.94±0.99
	Dark	94.95±0.53	0.26±0.43	57	0.150±0.052	[0.096, 0.235]	97.81±0.42
	Overexposure	94.94±0.03	0.33±0.26	105	0.277±0.070	[0.199, 0.385]	99.14±0.67
	Low contrast	95.89±0.16	0.21±0.31	65	0.172±0.033	[0.113, 0.260]	98.69±0.13
	Rotation	93.78±0.33	0.18±0.29	129	0.340±0.055	[0.253, 0.458]	96.86±0.59
	Occlusion	42.38±2.44	4.74±6.23	103	0.272±0.092	[0.195, 0.379]	75.00±3.03
EfficientNet-Lite0	Clean	97.73±0.19	0.08±0.11	20	0.053±0.024	[0.025, 0.111]	97.28±0.13
	Blur	68.05±1.31	2.83±2.52	324	0.855±0.508	[0.709, 1.031]	81.17±0.14
	Dark	96.51±0.31	0.19±0.21	12	0.032±0.021	[0.012, 0.081]	96.64±0.24
	Overexposure	95.25±0.15	0.28±0.24	55	0.145±0.030	[0.092, 0.228]	97.25±0.50
	Low contrast	97.06±0.31	0.13±0.17	44	0.116±0.057	[0.070, 0.193]	96.83±0.17
	Rotation	94.46±0.26	0.20±0.30	74	0.195±0.056	[0.132, 0.289]	96.14±0.21
	Occlusion	48.62±1.43	7.66±9.17	72	0.190±0.107	[0.128, 0.283]	79.28±1.19
ResNet18	Clean	98.00±0.15	0.71±0.43	43	0.113±0.040	[0.068, 0.189]	98.58±0.17
	Blur	68.54±1.28	11.79±3.77	122	0.322±0.171	[0.237, 0.437]	80.42±0.58
	Dark	97.04±0.12	1.53±0.48	20	0.053±0.032	[0.025, 0.111]	96.67±0.38
	Overexposure	95.05±0.24	1.47±0.88	89	0.235±0.088	[0.164, 0.336]	97.61±0.24
	Low contrast	97.57±0.16	1.15±0.58	21	0.055±0.029	[0.027, 0.114]	97.72±0.47
	Rotation	94.08±0.25	1.68±0.99	70	0.185±0.076	[0.123, 0.276]	96.50±0.44
	Occlusion	39.21±1.37	26.62±11.01	38	0.100±0.056	[0.058, 0.173]	72.00±1.28

Table A5 Complementary C_1 risk measures for the other three backbones (test set, three seeds pooled). Per frame: wrong C_1 cues divided by all test frames (the primary metric). Among spoken: wrong C_1 cues divided by all accepted cues whose predicted category is in C_1. Missed: protected frames not accepted with the correct category, divided by the 3 600 protected frames (1 200 per seed). Accepted misclassified: frames accepted with a wrong fine class, as a fraction of all frames (mean ± SD over seeds); for Speed signs the spoken phrase names the fine class, so this is an upper bound on any incorrect spoken content. Brackets: seed-adjusted 95% Wilson intervals

Backbone	Condition	Wrong per frame (%)	Wrong/spoken protected cues	Wrong among spoken (%)	Missed protected signs (%)	Accepted misclassified (%)
MobileNetV3-small	Clean	0.164	62/3 619	1.71 [1.12, 2.61]	1.19 [0.72, 1.98]	2.64±0.07
	Blur	0.929	352/3 194	11.02 [9.28, 13.04]	21.06 [18.84, 23.45]	30.54±2.54
	Dark	0.150	57/3 578	1.59 [1.02, 2.47]	2.19 [1.51, 3.19]	4.89±0.66
	Overexposure	0.277	105/3 674	2.86 [2.06, 3.95]	0.86 [0.47, 1.56]	4.85±0.16
	Low contrast	0.172	65/3 618	1.80 [1.19, 2.71]	1.31 [0.80, 2.12]	3.96±0.32
	Rotation	0.340	129/3 616	3.57 [2.66, 4.77]	3.14 [2.29, 4.28]	6.10±0.21
	Occlusion	0.272	103/2 803	3.67 [2.65, 5.08]	25.00 [22.63, 27.53]	53.83±4.60
EfficientNet-Lite0	Clean	0.053	20/3 522	0.57 [0.27, 1.19]	2.72 [1.94, 3.80]	2.22±0.26
	Blur	0.855	324/3 246	9.98 [8.33, 11.91]	18.83 [16.72, 21.14]	29.70±3.32
	Dark	0.032	12/3 491	0.34 [0.13, 0.88]	3.36 [2.48, 4.54]	3.37±0.42
	Overexposure	0.145	55/3 556	1.55 [0.98, 2.42]	2.75 [1.96, 3.84]	4.52±0.17
	Low contrast	0.116	44/3 530	1.25 [0.75, 2.06]	3.17 [2.32, 4.32]	2.85±0.39
	Rotation	0.195	74/3 535	2.09 [1.42, 3.08]	3.86 [2.91, 5.11]	5.41±0.37
	Occlusion	0.190	72/2 926	2.46 [1.66, 3.64]	20.72 [18.52, 23.11]	44.85±6.93
ResNet18	Clean	0.113	43/3 592	1.20 [0.72, 1.99]	1.42 [0.89, 2.26]	1.65±0.32
	Blur	0.322	122/3 017	4.04 [2.99, 5.45]	19.58 [17.44, 21.92]	21.83±3.79
	Dark	0.053	20/3 500	0.57 [0.27, 1.19]	3.33 [2.46, 4.51]	2.09±0.34
	Overexposure	0.235	89/3 603	2.47 [1.73, 3.51]	2.39 [1.66, 3.42]	3.97±0.71
	Low contrast	0.055	21/3 539	0.59 [0.29, 1.22]	2.28 [1.57, 3.29]	1.79±0.51
	Rotation	0.185	70/3 544	1.98 [1.32, 2.94]	3.50 [2.60, 4.70]	4.86±0.67
	Occlusion	0.100	38/2 630	1.44 [0.84, 2.47]	28.00 [25.53, 30.61]	38.88±7.65

[15] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778, https://doi.org/10.1109/CVPR.2016.90

[16] Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations

[17] Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations

[18] Howard A, Sandler M, Chu G, et al (2019) Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1314–1324, https://doi.org/10.1109/ICCV.2019.00140

[19] Kish L (1965) Survey Sampling. John Wiley & Sons, New York

[20] Kuriakose B, Shrestha R, Sandnes FE (2022) Tools and technologies for blind and visually impaired navigation support: a review. IETE Technical Review 39(1):3–18. https://doi.org/10.1080/02564602.2020.1819893

[21] Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30, pp 6402–6413

[22] Liu W, Wang X, Owens J, et al (2020) Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems 33, pp 21464–21475

[23] Loshchilov I, Hutter F (2017) SGDR: stochastic gradient descent with warm restarts. In: International Conference on Learning Representations

[24] Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International Conference on Learning Representations

[25] Manduchi R, Kurniawan S (2011) Mobility-related accidents experienced by people with visual impairment. Insight: Research and Practice in Visual Impairment and Blindness 4(2):44–54

[26] Müller R, Kornblith S, Hinton GE (2019) When does label smoothing help? In: Advances in Neural Information Processing Systems 32, pp 4694–4703

[27] Müller SG, Hutter F (2021) TrivialAugment: tuning-free yet state-of-the-art data augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 754–762, https://doi.org/10.1109/ICCV48922.2021.00081

[28] Naz I, Shah JH, Tahir A, et al (2025) Attention to detail: a conditional multi-head transformer for traffic sign recognition. PLoS ONE 20(11):e0335341. https://doi.org/10.1371/journal.pone.0335341

[29] Paszke A, Gross S, Massa F, et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp 8024–8035

[30] Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30(4):838–855. https://doi.org/10.1137/0330046

[31] Prandi C, Barricelli BR, Mirri S, et al (2023) Accessible wayfinding and navigation: a systematic mapping study. Universal Access in the Information Society 22(1):185–212. https://doi.org/10.1007/s10209-021-00843-x

[32] Russakovsky O, Deng J, Su H, et al (2015) ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

[33] Stallkamp J, Schlipsing M, Salmen J, et al (2012) Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition. Neural Networks 32:323–332. https://doi.org/10.1016/j.neunet.2012.02.016

[34] Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the Inception architecture for computer vision. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, pp 2818–2826, https://doi.org/10.1109/CVPR.2016.308

[35] Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp 6105–6114

[36] Tang W, Liu De, Zhao X, et al (2023) A dataset for the recognition of obstacles on blind sidewalk. Universal Access in the Information Society 22(1):69–82. https://doi.org/10.1007/s10209-021-00837-9

[37] Tian S, Zheng M, Zou W, et al (2021) Dynamic crosswalk scene understanding for the visually impaired. IEEE Transactions on Neural Systems and Rehabilitation Engineering 29:1478–1486. https://doi.org/10.1109/TNSRE.2021.3096379

[38] Valipoor MM, de Antonio A (2023) Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping. Universal Access in the Information Society 22(3):983–1005. https://doi.org/10.1007/s10209-022-00868-w

[39] Wang DB, Li L, Zhao P, et al (2023) On the pitfall of mixup for uncertainty calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7609–7618, https://doi.org/10.1109/CVPR52729.2023.00735

[40] Wightman R (2019) PyTorch image models. GitHub repository, https://github.com/huggingface/pytorch-image-models, https://doi.org/10.5281/zenodo.4414861

[41] Wilson EB (1927) Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953

[42] Wong A, Shafiee MJ, St. Jules M (2018) MicronNet: a highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification. IEEE Access 6:59803–59810. https://doi.org/10.1109/ACCESS.2018.2873948

[43] Yu S, Lee H, Kim J (2019) LYTNet: a convolutional neural network for real-time pedestrian traffic lights and zebra crossing recognition for the visually impaired. In: Vento M, Percannella G (eds) Computer Analysis of Images and Patterns (CAIP 2019), Lecture Notes in Computer Science, vol 11678. Springer, Cham, pp 259–270, https://doi.org/10.1007/978-3-030-29888-3_21

[44] Zhang H, Cisse M, Dauphin YN, et al (2018) mixup: beyond empirical risk minimization. In: International Conference on Learning Representations

[45] Zhong Z, Zheng L, Kang G, et al (2020) Random erasing data augmentation. Proceedings of the AAAI Conference on Artificial Intelligence 34(7):13001–13008. https://doi.org/10.1609/aaai.v34i07.7000

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip