International Journal of Graphic Design (IJGD)
Volume 3 No. 2 October 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 455-473
DOI: https://doi.org/10.51903/ijgd.v3i2.3991
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Computer-Vision-Informed Visual Explanation Cards for Autonomous-Driving Traffic-Sign Alerts: Localization, Classification, and Retrieved Evidence on GTSDB

Ruiyan Ma*[1], Long Zhang[2], Tiffany Song[3]
[1]Software Engineering, UC Irvine, CA, USA
[2]Transportation Systems Engineering, Southern Methodist University, TX, USA
[3]Human-Computer Interaction Design, Indiana University Bloomington, Bloomington, IN, USA
Email Address: maruiyan17@gmail.com

Abstract. *This paper develops a computer-vision-informed visual explanation card for traffic-sign alerts in autonomous-driving and driver-assistance interfaces. The card organizes a detected sign crop, predicted class, confidence, default semantic display-priority tier, retrieved visual precedents, scene-location cue, and concise action prompt. The empirical study uses the complete German Traffic Sign Detection Benchmark (GTSDB), comprising 900 road scenes in the standard 600-scene development and 300-scene evaluation portions. The same scenes support localization, crop classification, retrieval, calibration, and end-to-end analysis. The first 600 scenes were divided at the scene level into training and validation subsets; the 300 evaluation scenes were held out until model choices, retrieval depth, fusion weight, and detector threshold had been fixed. A learned class-agnostic localizer filters color-connected-component proposals with a histogram-of-oriented-gradients and color classifier. Five crop classifiers and four nearest-neighbor settings were evaluated, with retrieval treated primarily as example-based explanation support. On the held-out scenes, the selected localizer achieved an AP at IoU 0.50 of 0.211, an AP averaged over IoU 0.50–0.95 of 0.098, and a recall of 0.260 at the validation-selected operating point. The selected crop classifier achieved 0.784 accuracy and 0.574 macro-F1 on ground-truth crops. With predicted crops, correct-class end-to-end coverage was 0.177, and correct-tier end-to-end coverage was 0.244. These results define the information that the proposed card can receive from the evaluated vision pipeline. They do not measure driver comprehension, glance behavior, response time, trust, usability, or deployment safety, which require separate human-centred evaluation.*

Keywords: *traffic-sign detection; autonomous driving; computer vision; visual explanation; graphic design.*

INTRODUCTION

Traffic signs are a direct point of contact between machine perception and graphic design. Their colors, shapes, pictograms, and text encode road rules that must remain discriminable under motion, clutter, distance, changing illumination, and partial occlusion. In an autonomous-driving or advanced driver-assistance system, a camera may detect and classify a sign before the driver or remote operator has consciously interpreted it. The interface must then translate a model output into a compact visual message without concealing uncertainty or implying more capability than the perception system has demonstrated.

Recognition accuracy alone does not determine whether that message is well formed. Warning design depends on salience, timing, location, message economy, and a clear distinction between the event and the response it requests (Wickens et al., 2004; Wogalter, 2006). Situation-awareness theory similarly distinguishes perception of an element from comprehension of its meaning and projection of its implications (Endsley, 1995). For a traffic-sign alert, these

Received: August 2025; Revised: September 2025; Accepted: October 2025; Published: October 2025
*Corresponding author, maruiyan17@gmail.com

distinctions suggest a card that preserves the visible sign evidence, identifies the inferred class, communicates confidence, and gives the event an appropriate position in the display hierarchy.

The present study treats the alert card as the communication layer of a measured computer-vision pipeline. Its primary visual evidence is the detected sign crop. The predicted label and confidence describe the classifier output. A default semantic display-priority tier provides a static design prior based on the sign category, while retrieved training examples appear as compact visual precedents (Jin, 2025b). Retrieval is not assumed to be the principal source of recognition performance. Instead, it makes nearby cases in the selected feature space visible and allows the card to expose part of the local evidence surrounding a prediction. Such examples can support inspection, but they are neither causal explanations nor guarantees of correctness (Chen et al., 2019; Kim et al., 2016).

The study makes four contributions. First, it evaluates localization, crop classification, nearest-neighbor retrieval, calibration, and predicted-crop end-to-end coverage on the complete 900-scene GTSDB. Second, it compares an untrained color-proposal baseline with a learned proposal filter whose operating point is selected on validation scenes and then fixed for the held-out evaluation scenes. Third, it defines a three-level default semantic display-priority mapping and states the contextual variables that the dataset does not contain. Fourth, it translates the measured outputs and errors into a seven-part visual card specification.

The scope is deliberately bounded. The study asks whether the evaluated vision pipeline can supply coherent fields to a proposed traffic-sign card and where information is lost between scene localization and card population. It does not test whether drivers, remote operators, or HMI reviewers understand the card, look at it efficiently, respond appropriately, or calibrate reliance on the system (Chen & Xu, 2026). Findings from prior human studies motivate the need for explanation and appropriate reliance, but they do not establish the effectiveness of this particular interface (Koo et al., 2015; Lee & See, 2004).

LITERATURE REVIEW

Traffic-sign detection in full road scenes combines small-object localization with fine-grained category recognition (Xin, 2026). GTSDB formalized this problem with 900 images collected from driving sequences in Germany and a standard separation between 600 development scenes and 300 evaluation scenes (Houben et al., 2013). The signs occupy a small fraction of each 1360 × 800 image, and some scenes contain no annotated sign while others contain several. These properties make the benchmark useful for distinguishing three questions that are often blurred in interface demonstrations: whether a sign region is found, whether a found crop is classified correctly, and whether the resulting information can populate a card.

Engineered features remain useful for a compact, auditable baseline. Histograms of oriented gradients represent local edge orientation and have long supported object-recognition systems (Dalal & Triggs, 2005). Support-vector machines model high-dimensional class boundaries (Cortes & Vapnik, 1995), while random forests offer a nonlinear ensemble alternative (Breiman, 2001). The present study compares these model families with logistic regression and distance-weighted nearest neighbours. The aim is not to establish a new detection benchmark record; it is to obtain a controlled set of measured outputs from one dataset and to trace their consequences through the card pipeline.

Example-based explanations provide a complementary view of a prediction. Case-based reasoning explains a new instance by reference to related prior instances (Aamodt & Plaza, 1994). In image recognition, prototype-oriented approaches have shown how visual similarity can be exposed as part of an interpretable prediction (Chen et al., 2019). However, a handful of examples cannot fully describe a model or the underlying data distribution (Kim et al., 2016). Accordingly, the retrieved images in this study are called visual precedents or evidence chips. They show nearest training crops under a defined feature-space similarity, not the causal mechanism of the classifier and not proof that the predicted label is correct.

Confidence is a separate communication problem. A probability-like score can be visually persuasive even when it is poorly aligned with empirical correctness (Xin, 2025). Calibration measures such as multiclass Brier score, negative log-likelihood, and expected calibration error help reveal that mismatch (Guo et al., 2017; Zhou, Jin, & Zhao, 2025). The card therefore keeps confidence separate from display priority. Confidence describes the model's certainty about its own class output; priority represents a category-level design prior. Neither variable is treated as a substitute for the other.

The display-priority mapping draws on guidance for driver–vehicle messages rather than on an inferred accident-risk model. NHTSA guidance identifies safety relevance, operational relevance, potential consequence, and time available for response as important considerations, while also emphasizing that priority can change with the driving context (Campbell et al., 2016). ISO/TS 16951:2021 likewise concerns procedures for determining the priority of on-board messages without prescribing a complete interface or validating a particular message treatment (International Organization for Standardization [ISO], 2021). Since GTSDB does not contain vehicle speed, time-to-sign, lane or route relevance, traffic state, road condition, or automation mode, the class-level mapping used here is a default semantic prior rather than a measurement of current operational urgency (Zhou, Li, & Liu, 2025).

Graphic and interaction design determine how these machine outputs are arranged. Visibility, feedback, and mapping support understandable action possibilities (Norman, 2013;

Mendez & Okafor, 2026). Task-oriented visual encoding requires that color, position, and grouping correspond to the meaning of the data rather than act as decoration (Chen & Li, 2025; Munzner, 2014; Ware, 2012). Human-AI interaction guidance also recommends making system status and limitations legible (Amershi et al., 2019; Xu et al., 2025). These principles motivate a crop-first card in which the model output remains visibly connected to the road evidence, while confidence, display priority, and retrieved examples occupy distinct fields.

Traffic-sign comprehension research provides a further constraint: shape, symbol familiarity, and perceptual discriminability influence how road signs are interpreted (Ben-Bassat & Shinar, 2006). The interface should therefore preserve the sign image rather than replace it with a bare textual label. At the same time, a crop and a few nearby examples do not establish comprehension. Local explanation methods can expose evidence around an individual prediction (Ribeiro et al., 2016), but the effect of a particular card on people remains an empirical HMI question (Li et al., 2025).

METHODS

Dataset and evaluation protocol

The study used the complete GTSDB as its only image dataset. The benchmark contains 900 road-scene images at 1360 × 800 pixels. Scene identifiers 00000–00599 form the original development portion, and identifiers 00600–00899 form the original evaluation portion (Houben et al., 2013). The benchmark ground-truth file aligned to these scenes contains 1,213 annotated signs: 852 in the first 600 scenes and 361 in the final 300 scenes. All reported metrics use those benchmark boxes and class identifiers.

Table 1. GTSDB dataset overview and empirical role

Portion	Scene identifiers	Scenes	Annotated signs	Role
Training subset	Within 00000–00599	481	680	Fit detector, crop classifiers, feature scaler, and retrieval index
Validation subset	Within 00000–00599	119	172	Select model configuration, retrieval depth, fusion weight, and detector operating point
Held-out evaluation portion	00600–00899	300	361	Final localization, crop-classification, retrieval, calibration, and end-to-end evaluation
Complete benchmark	00000–00899	900	1,213	Source for the visual-card study

The first 600 scenes were split at the scene level into 481 training scenes and 119 validation scenes. Positive and negative scenes were both represented, multi-sign scenes remained intact, and any singleton-class scene needed to preserve all 43 classes in training was retained in the training subset. This produced 680 ground-truth training crops and 172 ground-truth validation crops. The 300 evaluation scenes and their 361 annotations were not used to choose a classifier,

hyperparameter, retrieval depth, fusion weight, confidence threshold, or non-maximum-suppression setting. Table 1 summarizes the data roles. An image-level integrity check found one exact duplicate pair within the 300-scene evaluation portion. The primary analysis retains the standard benchmark portion. A sensitivity analysis removes one member of that pair and repeats the localization metrics; the resulting difference is reported with the detector results rather than used to alter the model.

Crop representation and classifiers

Each annotated sign box was padded by 8%, converted to a square crop, and resized to 32 × 32 pixels. The descriptor combined a small histogram of oriented gradients with color and gradient summaries. HOG used 8 × 8-pixel cells, nine unsigned orientation bins, 2 × 2-cell blocks, and block normalization. Each RGB channel contributed its mean, standard deviation, and an eight-bin histogram. Global grayscale, gradient-magnitude, saturation, centre-region, and border-region summaries completed a 366-dimensional descriptor.

Five classifier families were compared: distance-weighted k-nearest neighbours; a linear support-vector machine; a radial-basis-function support-vector machine; multinomial logistic regression; and a random forest. The candidate values were fixed before final testing: (k {1,3,5,7}) for k-nearest neighbours; (C {0.1,1,10}) for the linear SVM and logistic regression; (C {1,10}) for the RBF SVM; and either square-root or 0.5 feature subsampling for a 400-tree random forest. Each family's configuration was selected by validation accuracy. The overall classifier used by the card pipeline was likewise selected by validation accuracy, after which the held-out crops were evaluated once.

The crop metrics were top-1 accuracy, macro-F1 over the union of labels present in the reference targets or model predictions, balanced accuracy, top-3 accuracy, and classifier scoring time on precomputed descriptors. Multiclass Brier score, negative log-likelihood, and ten-bin expected calibration error were computed from the models' probability outputs. The implementation used Python 3.12.13, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0 on an AMD EPYC 9V74 (9 logical CPU threads) CPU without GPU acceleration (Pedregosa et al., 2011).

Retrieved visual precedents and fusion

Retrieval used the same 366-dimensional crop descriptor. Training features were standardized, L2-normalized, and compared with query crops by cosine similarity. For (k {1,3,5,7}), class probabilities were formed by similarity-weighted voting among the nearest training crops. The retrieval depth was selected by validation accuracy, and the selected value was fixed for the held-out evaluation crops. The visual card uses the closest retrieved crops as evidence chips. Each chip should show the crop, class name, and similarity value. These

neighbours expose visually nearby training cases and may help a reviewer identify whether the query sits in a coherent or mixed neighbourhood (Zhang & Zhang, 2025a). They are not presented as independent confirmations of the prediction (Jin, 2025a).

A late-fusion analysis combined the validation-selected discriminative classifier probabilities with the selected retrieval probabilities (Xin, 2025). Candidate classifier weights were 0, 0.25, 0.50, 0.75, and 1.00, with the complementary weight assigned to retrieval. The weight was selected by validation accuracy and then fixed for final testing. Fusion was included to measure whether retrieved-neighbour probabilities changed recognition or calibration for the selected classifier, not to define retrieval as the primary recognition model.

Default semantic display-priority tiers

Each of the 43 traffic-sign classes was assigned a default semantic display-priority tier. Tier 1 contains signs whose meaning normally signals an immediate conflict rule or nearby hazard. Tier 2 contains speed, right-of-way, access, and mandatory-direction rules that alter the vehicle's operational constraints. Tier 3 contains signs that end a restriction or communicate a lower-default-priority state transition. Table 2 covers every class through eleven semantic families and states the rationale that applies to each explicitly enumerated class ID.

Table 2. Default semantic display-priority mapping and rationale for all 43 GTSDB classes

Class IDs	Traffic-sign family	Tier	Default semantic rationale
13, 14, 17	Yield, stop, and no entry	1	Require an immediate conflict decision, stop response, or path rejection
18, 26	General caution and traffic signals	1	Signal an unspecified nearby hazard or upcoming controlled conflict point
19–21	Dangerous curves and double curve	1	Signal an immediate road-geometry hazard
22–25	Bumpy/slippery road, narrowing, and road work	1	Signal a nearby surface, width, or work-zone hazard
27–29	Pedestrians, children, and bicycles crossing	1	Signal possible vulnerable road users near or crossing the path
30, 31	Ice/snow and wild animals	1	Signal a nearby traction hazard or sudden crossing hazard
0–5, 7, 8	Speed limits from 20 to 120 km/h	2	Set an operational speed constraint whose urgency depends on the scene
9, 10, 15, 16	Passing and vehicle-access restrictions	2	Establish manoeuvre or access constraints
11, 12	Right-of-way and priority road	2	Establish or change right-of-way expectations
33–40	Mandatory direction and roundabout signs	2	Prescribe the permitted path or junction movement
6, 32, 41, 42	Restriction-ending signs	3	End a prior speed, passing, or combined restriction

The mapping uses potential consequence, time sensitivity, safety relevance, and operational relevance as semantic criteria (Campbell et al., 2016; ISO, 2021). It is not an accident-risk estimate and does not determine final message priority in a moving vehicle. Speed, distance or

time to the sign, lane and route relevance, traffic state, road condition, and automation mode can raise, lower, suppress, or supersede the default tier. Class predictions were mapped to tiers after classification. Display-tier agreement is the proportion for which the predicted and ground-truth classes share the same tier. Tier-specific recall reports the proportion of ground-truth Tier 1, Tier 2, and Tier 3 signs whose predicted class remains in the same tier. These metrics evaluate consistency with the assigned mapping; they do not validate that the mapping improves human performance.

Learned class-agnostic localization

The localization component begins with broad red, blue, and yellow color masks. Connected components at two connectivity scales are converted into square candidate boxes at three size scales. Proposals outside broad size and aspect-ratio bounds are removed, and near-duplicate proposals are suppressed. This proposal stage is retained as a transparent color-salience baseline. The learned localizer adds a binary proposal filter. Positive detector crops were generated from each training annotation with small scale and translation variants. Negative crops were drawn from high-prior proposals whose intersection-over-union with every ground-truth box was below 0.10. A class-balanced SGD logistic model was fitted to the same HOG and color representation used for sign crops. At inference, the learned probability contributed 85% of each proposal score and the color prior contributed 15%. Non-maximum suppression used an IoU threshold of 0.35, and no more than 100 proposals were retained per scene.

The confidence threshold was selected by maximum validation F1 at IoU 0.50, with precision and recall used only as deterministic tie breakers. The fixed threshold was then applied to the held-out scenes. Localization metrics were AP at IoU 0.50, AP averaged over IoU 0.50–0.95, precision, recall, F1, false positives per image, predictions per image, and latency. The averaged AP follows the multi-threshold evaluation principle used in modern object-detection benchmarks (Lin et al., 2014). Latency includes image decoding, proposal generation, feature extraction, model scoring, and suppression.

Oracle-crop and predicted-crop card evaluation

Two card-input conditions were evaluated. The oracle-crop condition uses the benchmark box for every held-out sign and isolates crop classification, retrieval, and tier mapping from localization. The predicted-crop condition first matches detector boxes to ground-truth signs at IoU 0.50 and then classifies the matched detector crops. Conditional class accuracy describes classification among detected objects. End-to-end correct-class coverage divides the number of detected-and-correctly-classified objects by all held-out objects. The analogous correct-tier coverage counts detected objects whose predicted class maps to the same default tier as the ground truth. Unmatched detector boxes are counted as false cards.

Visual-card construction and scope

The proposed card contains seven signals: detected crop, predicted class, default semantic display-priority tier, confidence, three retrieved visual precedents, scene-location cue, and one concise action prompt. The crop and class label form the primary group. Tier and confidence are adjacent but visually distinct. Retrieved examples form a secondary evidence row (Zhang & Zhang, 2025b). The scene cue links the card back to the source image, and the action prompt states the sign's conventional instruction in plain language. Table 3 summarizes the empirical components and their design roles, while Figure 1 shows the complete flow.

Table 3. Experimental components and visual-card role

Component	Implementation	Measured output	Card role
Scene localization	Color proposals with learned HOG/color filter	Box, detector score, latency	Supplies scene cue and candidate crop
Crop representation	32 × 32 HOG and color descriptor	366-dimensional feature vector	Shared basis for classification and retrieval
Class recognition	Validation-selected model among five families	Class probabilities and top labels	Supplies sign name and alternatives
Calibration assessment	Brier, NLL, and ten-bin ECE	Confidence–correctness alignment	Constrains confidence display
Visual retrieval	Validation-selected cosine-neighbour depth	Crops, classes, similarities	Supplies example-based evidence chips
Display-priority mapping	Static class-to-tier map	Tier and tier-agreement metrics	Supplies default hierarchy prior
End-to-end analysis	Predicted box → crop → classifier → tier	Coverage and false-card rates	Shows information retained at card input

Figure 1 places retrieval beside, rather than before, the classifier: both receive the sign descriptor, but the classifier supplies the primary prediction and retrieval supplies visual precedents. No driver, remote-operator, or HMI-reviewer study was conducted. Consequently, the visual specification is assessed for technical coherence with the measured pipeline, not for comprehension, glance efficiency, response time, workload, trust, appropriate reliance, or safety.

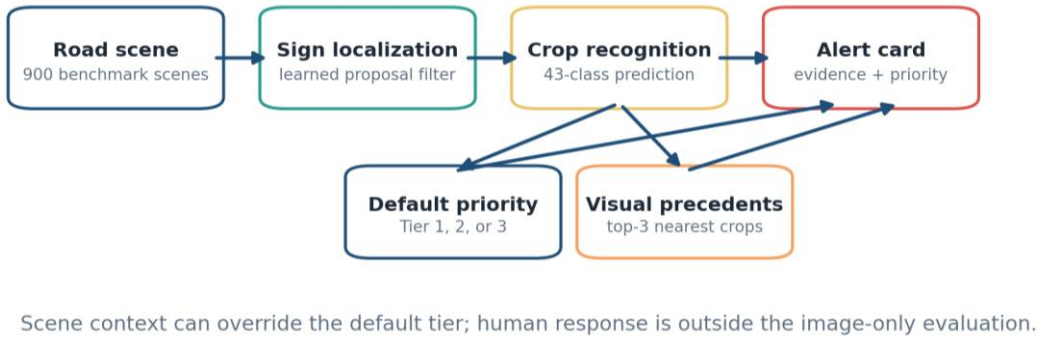

Figure 1. Computer-vision and visual-card pipeline evaluated on GTSDB.

RESULTS

Dataset distribution

The final scene and object counts are reported in Table 4. The training and validation split preserved scene boundaries and left all 300 benchmark evaluation scenes untouched. The evaluation portion contained 235 positive scenes, 65 negative scenes, and 361 annotated signs. Figure 2 shows the ground-truth crop distribution across the 43 classes for the training, validation, and evaluation portions. The long-tailed distribution makes macro-F1 and balanced accuracy important companions to overall accuracy.

Table 4. Scene and annotation statistics for the final GTSDB split

Split	Scenes	Positive scenes	Negative scenes	Annotated signs	Mean signs per scene	Classes represented
Training	481	406	75	680	1.414	43
Validation	119	100	19	172	1.445	36
Evaluation	300	235	65	361	1.203	38

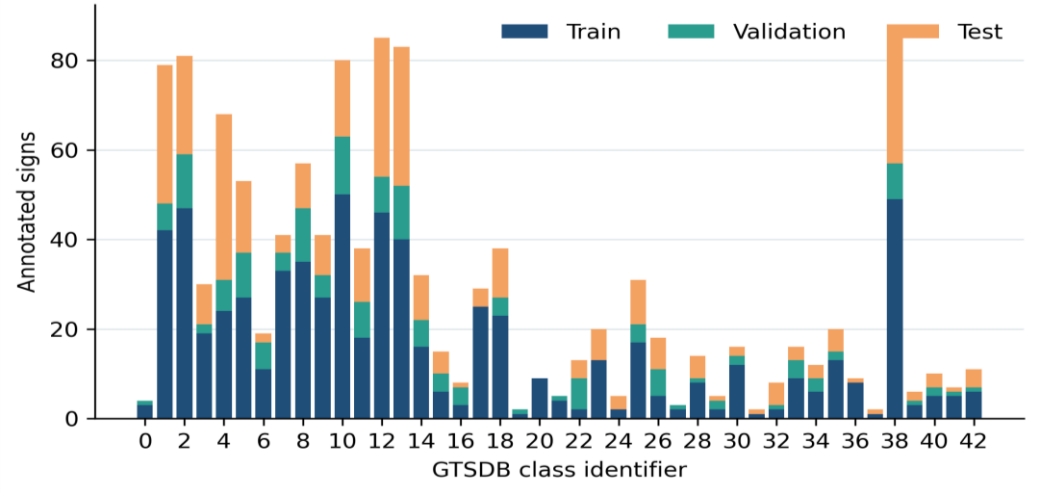

Figure 2. GTSDB class distribution for training, validation, and held-out evaluation crops.

Crop classification

Table 5 reports the configuration selected within each classifier family and its held-out crop results. Logistic regression with C=0.1 had the highest validation accuracy and was therefore fixed as the card pipeline's classifier. On 361 ground-truth evaluation crops, it achieved 0.784 top-1 accuracy, 0.574 macro-F1, 0.931 top-3 accuracy, and 0.636 balanced accuracy. Its mean classifier scoring time on the precomputed descriptors was 0.023 ms per crop on the reported CPU.

Table 5. Ground-truth-crop classification performance on the held-out GTSDB scenes

Classifier family	Validation-selected configuration	Accuracy	Macro-F1	Balanced accuracy	Top-3 accuracy	Descriptor scoring (ms/crop)

Distance-weighted kNN	k=7	0.651	0.419	0.448	0.878	0.027
Linear SVM	C=10.0	0.765	0.514	0.544	0.911	0.088
RBF SVM	C=10.0	0.789	0.506	0.543	0.895	0.164
Logistic regression	C=0.1	0.784	0.574	0.636	0.931	0.023
Random forest	max_features=sqrt	0.792	0.549	0.588	0.920	0.325

Figure 3 compares top-1 accuracy for the five discriminative classifiers with the validation-selected retrieval-only and fusion settings. The figure makes the roles visually explicit: the selected discriminative model is the primary recognizer, whereas retrieval is evaluated as a neighbour-based evidence mechanism and as a limited probability-fusion component.

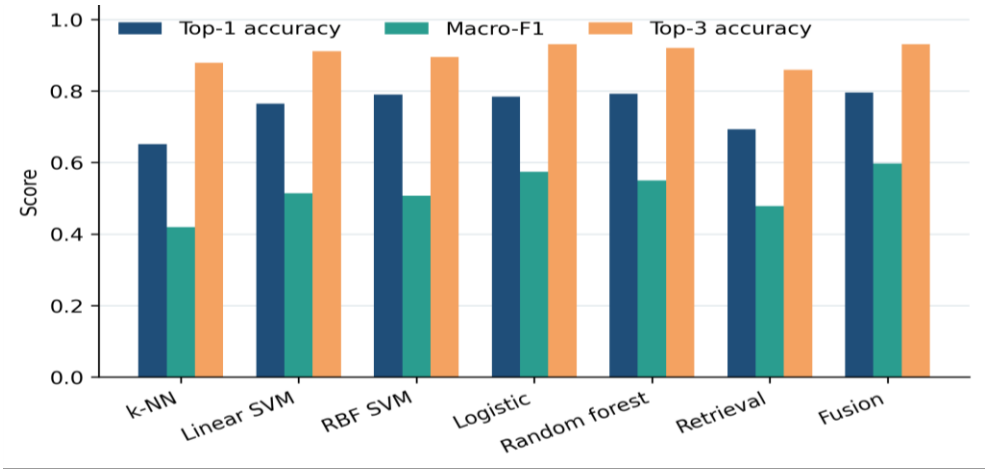

Figure 3. Held-out crop accuracy for classifiers, retrieval-only voting, and selected fusion.

Table 6. Retrieval-only and probability-fusion results on held-out ground-truth crops

Setting	Validation-selected parameters	Accuracy	Macro-F1	Top-3 accuracy	Top-1 neighbour same-class rate
Retrieval only	(k=5)	0.693	0.478	0.859	0.659
Fusion base classifier	logistic regression, C=0.1	0.784	0.574	0.931	—
Selected classifier–retrieval fusion	(w_c=0.750, w_r=0.250)	0.795	0.598	0.931	0.659

Retrieved evidence and fusion

Validation selected (k=5) for similarity-weighted retrieval. As shown in Table 6, retrieval-only classification achieved 0.693 accuracy and 0.478 macro-F1. The nearest retrieved crop had the same class as the query in 0.659 of evaluation cases. This rate describes local feature-space agreement; it does not establish that users would interpret the neighbours correctly. The selected fusion assigned 0.750 weight to the logistic regression classifier and 0.250 to retrieval. It achieved 0.795 accuracy, compared with 0.784 for its classifier-only counterpart, a difference of +0.011.

The result is reported as a model- and split-specific comparison rather than as a general performance claim for retrieval.

Default display-tier agreement

Table 7 reports agreement with the class-to-tier mapping. The selected classifier reached 0.953 overall tier agreement, with recalls of 0.938, 0.960, and 0.917 for Tiers 1, 2, and 3, respectively. Retrieval-only and fusion results are reported for comparison, but the tier metric remains conditional on the authored mapping.

Table 7. Agreement with the default semantic display-priority mapping

Setting	Overall tier agreement	Tier 1 recall	Tier 2 recall	Tier 3 recall
Selected classifier	0.953	0.938	0.960	0.917
Retrieval only	0.936	0.875	0.964	0.833
Selected fusion	0.956	0.927	0.968	0.917

Figure 4 shows the selected classifier's tier confusion matrix. The largest off-diagonal count was 9 Tier 2 crops assigned to Tier 1. Because the tiers are a default semantic grouping, the matrix indicates whether class confusions cross the proposed display levels; it does not indicate that the system estimated real-world hazard correctly.

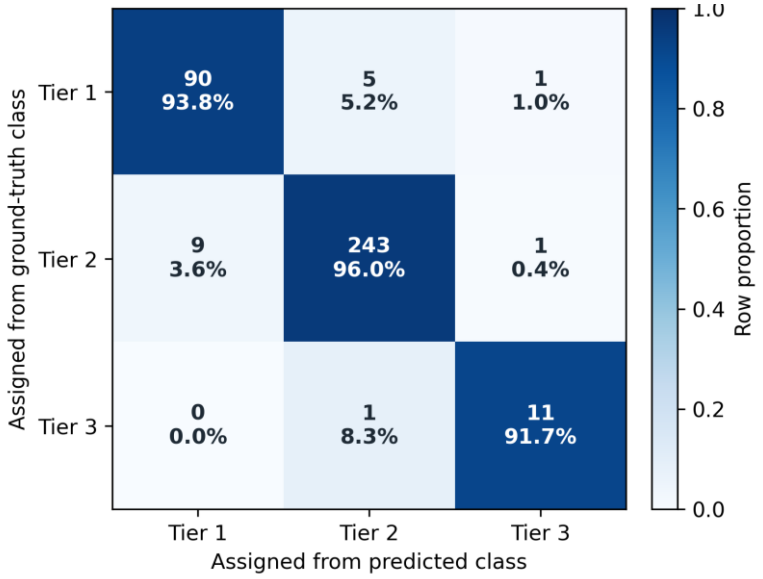

Figure 4. Default semantic display-priority confusion matrix for the selected crop classifier.

Road-scene localization

The learned proposal filter and the untrained color-proposal baseline are compared in Table 8. At the validation-selected operating point, the learned localizer achieved 0.211 AP50, 0.098 AP50–95, 0.495 precision, 0.260 recall, and 0.341 F1. It produced 0.320 false positives per scene

and required 509.4 ms per scene on average, including decoding and preprocessing. The color-only baseline achieved 0.026 AP50 and 0.068 F1.

Table 8. Class-agnostic localization on the held-out GTSDB scenes

Method	AP50	AP50–95	Precision	Recall	F1	False positives /image	Mean latency (ms/scene)
Color-connected-component baseline	0.026	0.015	0.062	0.075	0.068	1.350	274.9
Learned HOG/color proposal filter	0.211	0.098	0.495	0.260	0.341	0.320	509.4
Learned filter, duplicate-pair sensitivity	0.209	0.096	0.492	0.259	0.339	0.321	—

The duplicate-pair sensitivity analysis recomputed localization accuracy but did not repeat the timing benchmark. The learned filter changed the color baseline by +0.185 AP50 and +0.186 recall. Figure 5 illustrates a held-out road scene in which all four annotated signs were matched by validation-thresholded detector boxes. This representative success makes the detector output legible, while the aggregate missed signs and false detections reported in Table 8 show why localization performance must be carried into the card-level analysis rather than hidden behind ground-truth crops. The results characterize a compact learned baseline; they do not support deployment-readiness or safety claims.

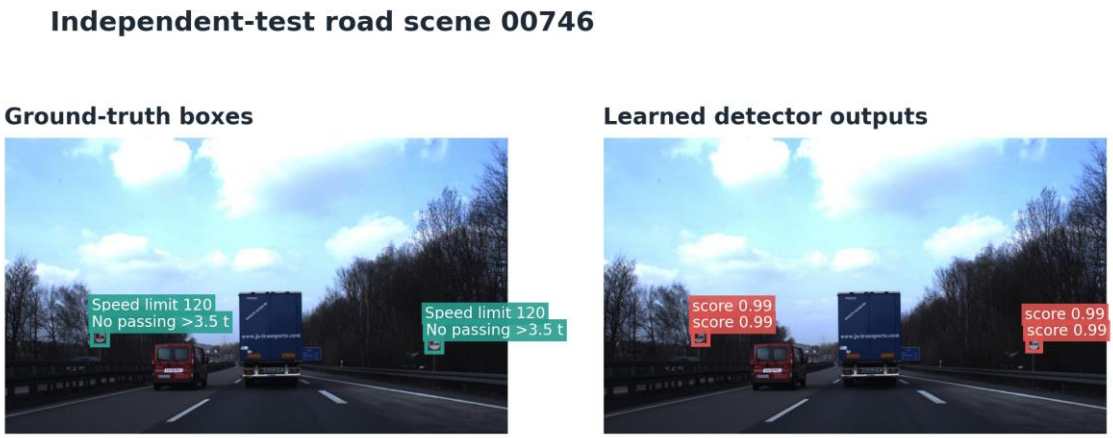

Figure 5. Held-out GTSDB scene with ground-truth and learned-detector boxes.

Calibration and card-input coverage

Calibration results for the selected classifier, retrieval-only probabilities, and fusion appear in Table 9. The selected classifier produced a multiclass Brier score of 0.316, an NLL of 0.787, and a ten-bin ECE of 0.106. The selected fusion produced 0.322, 0.800, and 0.131, respectively. These values determine how cautiously the numerical confidence field should be interpreted. They do not show that users will understand or appropriately rely on the displayed score.

Table 9. Calibration metrics for held-out ground-truth crops

Setting	Accuracy	Multiclass Brier	NLL	ECE (10 bins)
Selected classifier	0.784	0.316	0.787	0.106
Retrieval only	0.693	0.440	3.508	0.062
Selected fusion	0.795	0.322	0.800	0.131

Table 10 separates oracle-crop recognition from the predicted-crop pipeline. Oracle crops provide complete object coverage and isolate the classifier, yielding 0.784 class accuracy and 0.953 tier agreement. The detector matched 94 of 361 held-out objects, or 0.260. Among matched detector crops, class accuracy was 0.681 and tier agreement was 0.936. After missed signs were included in the denominator, correct-class coverage was 0.177 and correct-tier coverage was 0.244. Unmatched detections would have produced 0.320 false cards per scene.

Table 10. Oracle-crop and predicted-crop card-input performance

Condition	Ground-truth objects	Detected objects	Object coverage	Conditional class accuracy	Conditional tier agreement	End-to-end correct-class coverage	End-to-end correct-tier coverage	False cards/image
Ground-truth boxes (oracle crops)	361	361	1.000	0.784	0.953	0.784	0.953	0
Learned-detector boxes	361	94	0.260	0.681	0.936	0.177	0.244	0.320

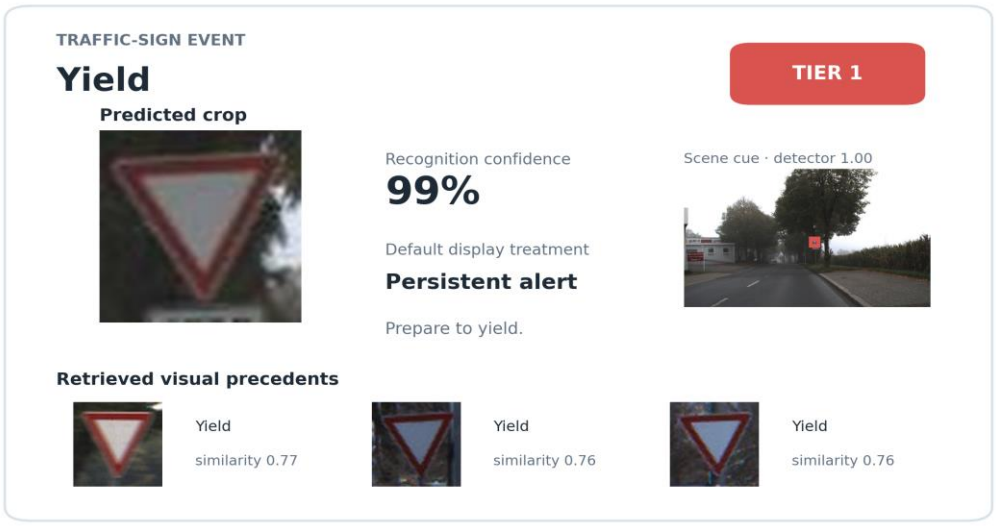

Figure 6. Visual explanation card populated by a held-out prediction and retrieved GTSDB precedents.

Figure 6 instantiates the card with one held-out predicted crop and its actual retrieved training neighbours. The detector box and sign crop preserve the link to the scene; the label and

confidence report the model output; the default tier controls visual prominence; and the evidence chips expose nearby examples with their similarities. The figure is a design specification populated by measured outputs, not evidence of human-interface effectiveness. Figure 7 gives a reliability diagram and confidence distribution for the selected classifier. The gap between bin accuracy and mean confidence visualizes the calibration quantities in Table 9 and supports using both a numerical value and a restrained verbal uncertainty label.

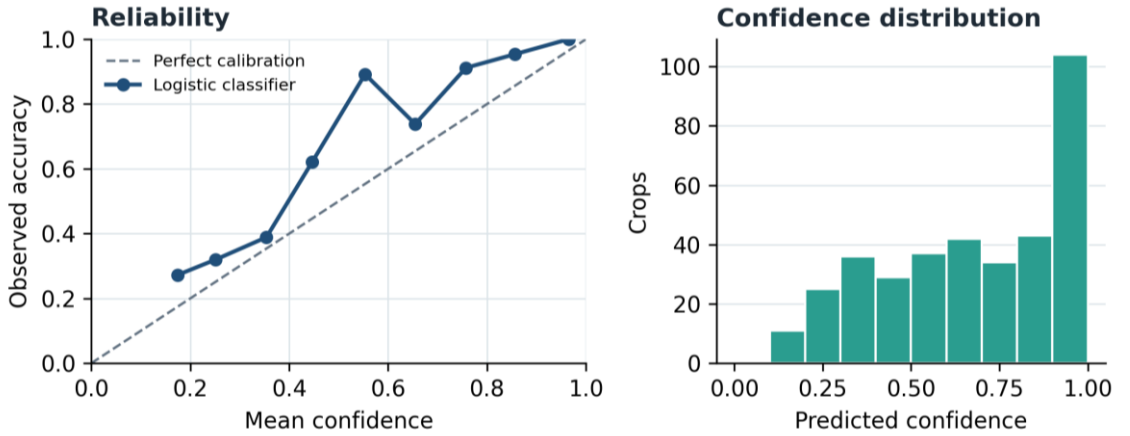

Figure 7. Reliability and confidence distribution of the selected classifier on held-out GTSDB crops.

Table 11. Visual encoding specification and interpretation boundaries

Data signal	Visual encoding	Empirical basis	Interpretation boundary
Detected sign crop	Largest image in the card, adjacent to the label	Predicted box and crop	May be absent or mislocalized when detector coverage fails
Predicted class	Plain-language name with sign pictogram	Selected classifier output	Can be wrong even when the detector box is correct
Default display-priority tier	Tier 1, 2, or 3 badge with position and restrained color contrast	Class-to-tier mapping and tier agreement	Static semantic prior; context may override it
Confidence	Numeric value plus calibrated verbal band	Probability and calibration metrics	Model certainty, not urgency or safety
Retrieved visual precedents	Three small crop chips with class and similarity	Nearest training crops	Similar examples, not causal proof or independent confirmation
Scene-location cue	Thin box in a road-scene thumbnail	Detector coordinates and score	Shows where the system looked; does not validate the detection
Action prompt	One short sign-specific instruction	Predicted sign semantics	Wording and timing require scenario and user evaluation

Visual encoding specification

Table 11 converts the measured pipeline into a visual specification. The design keeps observed evidence, model inference, static priority, and interface action in separate visual fields. This separation prevents a high confidence score from automatically becoming a high-priority

warning and prevents a Tier 1 category from being shown as though the model were certain. The specification follows a crop-first hierarchy because the road evidence remains directly inspectable. Tier, confidence, and retrieval are secondary encodings with different meanings. Their combined presence may support later evaluation of comprehension and appropriate reliance, but the current computer-vision experiment cannot determine those outcomes.

Per-class error pattern

Table 12 reports a prespecified set of class-level outcomes: the six lowest and six highest held-out class accuracies among classes with at least 5 evaluation examples. The fixed support rule prevents isolated one-example classes from dominating the comparison. The lowest-performing classes were concentrated in visually similar speed-limit and regulatory signs, whereas the highest-performing classes showed more distinctive shapes or well-represented regulatory symbols. These results support showing the observed crop and alternatives rather than a bare label, particularly for visually similar sign families.

Table 12. Selected held-out per-class outcomes from the validation-selected classifier

Selection group	Class ID	Sign class	Default tier	Test (n)	Class accuracy
Lowest 1	28	Children crossing	1	5	0.200
Lowest 2	32	End of all speed and passing limits	3	5	0.200
Lowest 3	26	Traffic signals	1	7	0.286
Lowest 4	8	Speed limit (120 km/h)	2	10	0.300
Lowest 5	11	Right-of-way at the next intersection	2	12	0.333
Lowest 6	25	Road work	1	10	0.500
Highest 6	35	Ahead only	2	5	1.000
Highest 5	15	No vehicles	2	5	1.000
Highest 4	14	Stop	1	10	1.000
Highest 3	10	No vehicles passing over 3.5 t	2	17	1.000
Highest 2	38	Keep right	2	31	1.000
Highest 1	13	Yield	1	31	1.000

CONCLUSION

This study connects traffic-sign perception to the visual structure of an autonomous-driving alert card using one complete road-scene benchmark. The final pipeline keeps localization, crop classification, retrieved precedents, confidence, default semantic display priority, and card population analytically separate. On held-out GTSDB scenes, the learned localizer achieved 0.211 AP50 and 0.260 recall, while the validation-selected classifier achieved 0.784 accuracy on ground-truth crops. When detector misses and localization errors were included, correct-class end-to-end coverage was 0.177 and correct-tier coverage was 0.244. The difference between

oracle-crop and predicted-crop results is the most direct measure of how localization constrains the card.

Retrieval is best understood as example-based explanation support. Its neighbours show which training crops are close to the query under the selected representation, and its probability contribution can be compared with the discriminative classifier through independently selected fusion. The held-out retrieval-only accuracy of 0.693 and fusion accuracy of 0.795 should be read alongside, not in place of, the selected classifier's performance. The card should label these images as similar visual precedents and show their similarity values rather than imply that they verify the prediction.

The three display tiers provide a consistent default hierarchy for a static card specification, but they are not a risk model. Their class-level assignment is based on sign semantics, potential consequence, time sensitivity, safety relevance, and operational relevance. Actual message priority requires variables that GTSDB does not provide, including vehicle speed, distance to the sign, lane and route relevance, traffic state, road condition, and automation mode. A deployed system would need a contextual priority layer above the static mapping.

The proposed visual grammar should therefore remain narrow. The sign crop and predicted label form the primary message. Confidence describes model certainty, while the tier describes default display priority; neither should be converted into the other. Retrieved examples remain secondary evidence, and the scene box shows the source of the crop. A concise action prompt may restate the sign's conventional instruction, but its wording, timing, persistence, and modality require evaluation in the intended operational setting.

The next empirical step is a human-centred study with drivers, remote operators, or HMI reviewers. Such a study should compare card variants using comprehension accuracy, glance duration, response time, workload, false-alarm response, and appropriate-reliance measures. It should also test context-dependent priority changes and card behaviour when localization fails, confidence is low, or retrieved examples disagree. Until those measures are available, the present contribution is a technically evaluated visual-card specification and an end-to-end account of the computer-vision information available to populate it.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(1), 39–59. https://doi.org/10.3233/AIC-1994-7104

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E.

(2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Ben-Bassat, T., & Shinar, D. (2006). Ergonomic guidelines for traffic sign design increase sign comprehension. Human Factors, 48(1), 182–195. https://doi.org/10.1518/001872006776412298

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campbell, J. L., Brown, J. L., Graving, J. S., Richard, C. M., Lichty, M. G., Sanquist, T., Bacon, L. P., Woods, R., Li, H., Williams, D. N., & Morgan, J. F. (2016). Human factors design guidance for driver-vehicle interfaces (Report No. DOT HS 812 360). National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/812360_humanfactorsdesignguidance.pdf

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. Advances in Neural Information Processing Systems, 32, 8930–8941. https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html

Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. Journal of Technology Informatics and Engineering, 4(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. International Journal of Graphic Design, 4(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 1, pp. 886–893). IEEE. https://doi.org/10.1109/CVPR.2005.177

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., & Igel, C. (2013). Detection of traffic signs in real-world images: The German Traffic Sign Detection

Benchmark. In 2013 International Joint Conference on Neural Networks (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN.2013.6706807

International Organization for Standardization. (2021). Road vehicles—Ergonomic aspects of transport information and control systems (TICS)—Procedures for determining priority of on-board messages presented to drivers (ISO/TS 16951:2021). https://www.iso.org/standard/81103.html

Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. Journal of Technology Informatics and Engineering, 4(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. International Journal of Graphic Design, 3(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. Advances in Neural Information Processing Systems, 29, 2280–2288. https://proceedings.neurips.cc/paper_files/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. International Journal on Interactive Design and Manufacturing, 9, 269–275. https://doi.org/10.1007/s12008-014-0227-2

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. International Journal of Graphic Design, 3(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), Computer vision—ECCV 2014 (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Mendez, L., & Okafor, S. (2026). Adaptive Graphic Interaction Model: A Mixed-Method Framework for Future Factory Design. International Journal of Graphic Design, 4(1), 1–16. https://doi.org/10.51903/IJGD.V4I1.3194

Munzner, T. (2014). Visualization analysis and design. CRC Press.

Norman, D. A. (2013). The design of everyday things (Rev. and expanded ed.). Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn:

Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Ware, C. (2012). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann.

Wickens, C. D., Lee, J. D., Liu, Y., & Gordon-Becker, S. E. (2004). An introduction to human factors engineering (2nd ed.). Pearson Prentice Hall.

Wogalter, M. S. (Ed.). (2006). Handbook of warnings. Lawrence Erlbaum Associates.

Xin, Q. (2025). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). Journal of Technology Informatics and Engineering, 4(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

Xin, Q. (2026). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. Journal of Technology Informatics and Engineering, 5(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. Journal of Technology Informatics and Engineering, 4(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. Journal of Technology Informatics and Engineering, 4(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. International Journal of Graphic Design, 3(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

Zhou, B., Jin, J., & Zhao, D. (2025). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. Journal of Technology Informatics and Engineering, 4(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. International Journal of Graphic Design, 3(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696