Handling Open-Set Recognition

Oriol Miró and Marc González

MAI, UPC
Deep Learning, Practice 2
June 12, 2025

Deep neural traffic-sign classifiers assume every test image belongs to a known class, yet deployed systems inevitably encounter unseen symbols. We address this open-set recognition challenge with a two-stage pipeline: an unsupervised convolutional auto-encoder first decides whether an input is in- or out-of-distribution, then forwards inliers to a fine-tuned ResNet-50 and outliers to a frozen CLIP zero-shot head. Evaluated on a head–tail split of the German GTSRB dataset and on a cross-dataset setting where Chinese TSRD signs are outliers, the same vanilla reconstruction auto-encoder yields an F1 of 0.75% and 0.95% for each split, respectively. A latent Mahalanobis filter helps only in the harder head–tail regime, confirming that Gaussian assumptions are context-dependent. The full system achieves 88.25% Top-1 accuracy and 61.92% F1 across 101 labels, showing that simple reconstruction AEs combined with off-the-shelf zero-shot vision–language models provide a strong baseline for safety-critical open-set recognition.

1 Introduction

Deep neural classifiers excel whenever test data follow *exactly* the same distribution as the training set. Unfortunately they can behave *arbitrarily* in the presence of truly novel inputs. Such brittle behaviour is unacceptable in safety-critical domains like autonomous driving, where misinterpreting an unseen traffic sign could have catastrophic consequences [2]. This mismatch between closed-world training and open-world deployment is formalised as the *open-set recognition* (OSR) problem.

OSR research has converged on three broad solution families. For example, with **out-of-distribution (OOD) detectors** a front-end gate tries to decide whether the input stems from the known data distribution; if so, the downstream classifier proceeds as usual, otherwise the system abstains or triggers a fall-back routine. In other approaches, such as **Open-set classifiers**, the network is retrained to emit an explicit "`unknown`" label, usually by adding a dummy class or by reshaping the loss landscape so that unknown samples are mapped to low confidence.

We adopt **Approach 1** and develop an *autoencoder OOD detector* (see Eq. 1. In it, an unsupervised convolutional auto-encoder (AE) is trained solely on in-distribution (ID) images. At test time the AE reconstruction error (which we also test combined with a Mahalanobis distance in latent space [1]) serves as the OOD score.

$$\text{image} \xrightarrow{\text{AE gate}} \begin{cases} \textbf{ID}: & \text{ResNet-50} \rightarrow \hat{y}_{\text{ID}} \\ \textbf{OOD}: & \text{CLIP (zero-shot)} \rightarrow \hat{y}_{\text{OOD}} \end{cases} \tag{1}$$

We implement and thoroughly evaluate a plain reconstruction AE gate and its Mahalanobis-augmented variant [1] under *two* regimes: (i) head–tail split *within* GTSRB, and (ii) cross-dataset detection between GTSRB (ID) and TSRD (OOD). We also attach a CLIP [6] zero-shot head to supply a *best-effort* label for images flagged as OOD, thereby satisfying the optional requirement of the assignment.

The remainder of this report is organised as follows: Section 2 details the two datasets; Section 3 describes the closed-set classifier, the auto-encoder detector and the zero-shot fallback; Section 4 reports quantitative and qualitative results; Section 5 reviews the results and discusses them; finally, Section 6 summarises our findings and outlines future work.

2 Datasets

We rely on two publicly-available datasets of cropped traffic-sign images. The long-established *German Traffic-Sign Recognition Benchmark* (**GTSRB**) [3] provides our in-distribution (ID) data, whereas the *Chinese Traffic-Sign Recognition Database* (**TSRD**) [4] supplies a foreign, but still road-sign so same-domain, OOD set.

Table 1 describes some features of the datasets, while Figures 1a and 1b show some examples from each dataset: on the left are German signs, and the right Chinese signs.

Table 1: *Key statistics of the two corpora.*

Dataset	Classes	Images (train / test)	Native resolution
GTSRB	43	39 209 / 12 630	15–250px
TSRD	58	4 170 / 1 994	variable, JPEG

(a) *Sample images from the GTSRB dataset [3].* **(b)** *Sample images from the TSRD dataset [4].*

Figure 1: *Examples from the German (left) and Chinese (right) traffic sign datasets.*

The reason for having *two* datasets is because we conduct two experimental rounds, as further detailed in §3. First, we stay *inside* GTSRB: the 25 most-frequent classes (*head*) train the models, while the 18 least-frequent classes (*tail*) are withheld and later presented as "unkowns", OOD. Should head–tail separation prove unreliable and too similar as we use the same dataset, we retrain on *all* 43 German classes and use TSRD as OOD. This second round keeps the task domain-consistent (traffic signs in both cases) yet introduces new glyph shapes, colour palettes and capture conditions, yielding a cleaner view of true distributional shift.

German Traffic-Sign Recognition Benchmark (GTSRB) Released with IJCNN-2011 [7], GTSRB totals 51 839 cropped RGB signs captured from dash-cam video in Germany. Images vary from 15×15 to 250×250 and cover 43 categories spread across regulatory, warning and priority groups. The official split allocates 39 209 images to training and 12 630 to testing. Because class frequencies are highly skewed, we label the 25 most common categories as *head* (ID) and hold the remaining 18 out as *tail* (in-dataset OOD).

Chinese Traffic-Sign Recognition Database (TSRD) TSRD has 6 164 cropped signs photographed on Chinese highways, partitioned into 4 170 training and 1 994 test images. All 58 categories map into the standard Chinese super-groups *prohibitory*, *mandatory* and *danger*. The Kaggle archive `traffic-sign-dataset-classification` publishes the original data and adds a ready-to-use `labels.csv` mapping[1].

3 Methodology

First, we ask whether an OOD detector trained and tuned *inside* the German Traffic-Sign Recognition Benchmark (GTSRB) [3] would work; that is, we use the *same* dataset, and split it into the 25 most frequent classes as the ID and the 18 least-frequent classes as OOD. If this proves too difficult and the reconstruction error still overlaps heavily—as indicated in preliminary experiments—we repeat the very same detector on a *foreign-sign* set, Chinese TSRD [4], where images stay within the traffic-sign domain yet diverge more strongly from the most-frequent classes. In this case, the full German GTSRB would be the ID and the full Chinese TSRD would be OOD. Moreover, we will also use a CLIP zero-shot model in the case we predict an image as OOD—but we will do so *only* on the best of the aforementioned approaches.

As a disclaimer to the reader, when performing OOD-detection we treat ID as the positive class. We feel the need to explicitly state this to avoid confusion, as after reading on the topic we found literature tended to treat OOD as the positive class.

The pipeline for the whole system was previosuly show in Eq. 1. The remainder of this section describes every component.

3.1 Closed-set Classifier

We fine-tune an **ImageNet-pre-trained ResNet-50**, chosen for its favourable accuracy-to-size ratio. On the first case, where we only use the same dataset, we fine-tune using the 25 head classes of GTSRB. On the second case, where GTSRB is the ID dataset and TSRD is the OOD dataset, we fine-tune on the whole GTSRB dataset. We follow standard transfer-learning recipes: for five warm-up epochs the convolutional backbone is frozen and only a freshly initialised $2048\to25$ linear head is optimised with SGD (momentum 0.9, LR 10^{-3}, weight-decay 10^{-4}). Afterwards we unfreeze the backbone, halve the learning rate, and keep training until the F_1 on the validation split stalls for eight epochs, at which point the LR is halved again ("reduce-on-plateau"). As in GTSRB images have different sizes (ranging from 15 to 250 pixels), we resize them to 224×224.

3.2 Auto-encoder OOD detector

Reconstruction-based OOD detectors assume that an auto-encoder trained on id images will fail to reproduce truly novel samples, thus yielding a larger error [2, 8]. Although this assumption can break for visually homogeneous data, it remains a natural first baseline and serves as the backbone for the Mahalanobis variant proposed by Denouden *et al.* [1].

Encoder–Decoder Architecture The auto-encoder is a *symmetric* CNN that compresses a 224×224 colour image into a compact latent vector and then restores that vector back to the original resolution.

Encoder. Starting from the input tensor $x \in [0,1]^{3\times224\times224}$, we have five convolutional blocks that execute the pattern:

$$\text{conv}_{4\times4,\ \text{stride}=2} \to \text{batch-norm} \to \text{ReLU}$$

[1] Available here: `https://www.kaggle.com/datasets/ahemateja19bec1025/traffic-sign-dataset-classification`

Each block halves the spatial dimensions while doubling the channel count, so the feature map sequence is:

$$\underbrace{(3, 224)}_{\text{input}} \;\to\; (32, 112) \;\to\; (64, 56) \;\to\; (128, 28)$$
$$\to\; (256, 14) \;\to\; (512, 7) = h_5,$$

where a pair (C, W) denotes "C channels, $W \times W$ pixels"; height equals width at all stages. We then flatten h_5 ($512 \times 7 \times 7 = 25\,088$ scalars) and feed it through a fully connected layer $W_e \in \mathbb{R}^{d \times 25\,088}$, yielding the latent code

$$z = W_e \,\mathrm{vec}(h_5) \in \mathbb{R}^d,$$

with $d \in \{2^2, 2^4, \ldots, 2^7\}$ explored in our ablations.

Decoder. The decoder first expands z back to a $512 \times 7 \times 7$ tensor and then *mirrors* the encoder using transposed convolutions ("deconvs") with the same kernel 4×4, stride 2 and batch-norm + ReLU:

$$512 \to 256 \to 128 \to 64 \to 32 \to 3.$$

Every step doubles the spatial size and halves the channel depth, exactly undoing the encoder's downsampling. At the end, we have a sigmoid that turn the output to $\hat{x} = g_\theta(z) \in [0, 1]^{3\times224\times224}$, matching the input dimension.

Autoencoder Training The auto-encoder is trained on ID images by minimising the per-pixel mean–squared error:

$$\mathcal{L}(\phi, \theta) = \frac{1}{CHW} \sum_{c,h,w} \big(x_{c,h,w} - \hat{x}_{c,h,w}\big)^2, \qquad (2)$$

Where C=3, H=W=224. We optimise with Adam (LR 10^{-3}), and the learning rate is also halved whenever there's no improvement for five epochs.

For every training image we record the mean squared error $s_{\text{rec}}(x) = \frac{1}{CHW}\|x - \hat{x}\|^2$. To keep the detector *fully unsupervised*, we set a single threshold $\tau^\star$ to the 95^{th} percentile of $\{s_{\text{rec}}(x_i)\}_{x_i \in \text{ID-val}}$, so at most 5 % of ID-validation images are ever rejected. We recompute this for every latent dimension; the resulting $\tau^\star$ is then reused *unchanged* at test time for all detector variants.

Autoencoder Variants We propose two variants, a *Vanilla* one, and one enhanced by the latent Mahalanobis distance. Both variants are evaluated with the same calibration grid for $\tau^\star$ and share the exact encoder–decoder architecture.

Variant 1: Vanilla Reconstruction error. Reject if $s_{\text{rec}}(x) > \tau^\star$.

Variant 2: Latent Mahalanobis distance. Some OOD images can still be reconstructed well if they lie on the latent manifold but far from the ID cluster. We therefore measure

$$d_M(x) = (z - \mu)^\top \Sigma^{-1} (z - \mu), \qquad (3)$$

where μ and Σ are the mean and covariance of latent codes for ID-train. Using the 95-th percentile τ_M of d_M as cut-off, our final rule is

$$\text{accept as ID} \iff \big[s_{\text{rec}}(x) \le \tau^\star\big] \wedge \big[d_M(x) \le \tau_M\big],$$

so we can combine the pixel-level and representation-level evidence in a single test.

3.3 Zero-shot fallback with CLIP

When the OOD gate rejects an image we still wish to provide a *best-effort* label rather than returning a bare "unknown". We therefore add a zero-shot classifier built on the public `OpenCLIP-ViT-B/32-LAION2B-s34B-b79K` checkpoint [6, 5]. We perform no training; all knowledge is transferred from the pre-trained model.

For every foreign TSRD class c we write a single, succinct English sentence that literally describes the sign, *e.g.*,

> "a photo of a dangerous left-curve traffic sign"
> "a photo of a …"

All prompts are encoded once and cached.

To perform the embedding matching, let g_t and g_i denote CLIP's frozen text and image encoders. We normalise their outputs to the unit sphere:

$$e_c = \frac{g_t(\text{prompt}_c)}{\|g_t(\text{prompt}_c)\|_2}, \qquad v(x) = \frac{g_i(x)}{\|g_i(x)\|_2}. \qquad (4)$$

Stacking the text vectors into $E = [e_1, \ldots, e_C] \in \mathbb{R}^{D\times C}$, inference requires only a dot-product lookup:

$$s(x) \;=\; E^\top v(x) \in \mathbb{R}^C, \qquad \hat{c}(x) = \arg\max_c s_c(x). \quad (5)$$

Because E is fixed, inference adds only one matrix–vector product, keeping latency negligible.

3.4 Evaluation protocol

Training relies solely on GTSRB. From the 39 209 images in the official training split we create an 80/20 stratified partition: the larger fold fits both the ResNet-50 classifier and every auto-encoder, whereas the smaller fold serves two purposes: learning-rate scheduling and calibration of the unsupervised thresholds used by the OOD gate. For each latent size the reconstruction-error cut-off is set to the 95^{th} percentile of the *ID-validation* error distribution and, when the Mahalanobis distance is employed, its cut-off is set in the same way. No OOD image is ever back-propagated through the networks.

Hyper-parameter search proceeds in two steps. First, the latent dimensionality d is chosen solely on the ID validation fold, Second, with that d frozen, we pick between the *Vanilla* and *Mahalanobis* auto-encoder variants by comparing their F_1 scores on the ID validation set *together with* a matching OOD validation set (either

the 18 German tail classes or the corresponding slice of TSRD).

Because the zero-shot CLIP component has no trainable parameters, we simply report its Top-1 accuracy and F_1 on the union of the OOD train and validation splits; no tuning is required.

Final testing evaluates the frozen pipeline on the untouched ID and the withheld OOD test split. We measure closed-set precision, recall and F_1; gate-only F_1 when the detector is fed mixed ID/OOD data; and end-to-end Top-1 accuracy and F_1 over the full 101-class label space (43 German + 58 Chinese). All experiments run on a BSC cluster.

4 Experiments

We study two complementary evaluation regimes, each trained from scratch with its *own* ResNet-50 classifier and autoencoder variants; only the code base carries over.

Head–Tail (intra-dataset)
ID = 25 most-frequent GTSRB classes
OOD = the 18 least-frequent GTSRB classes withheld from training (pilot set-up in§3).

Foreign-sign (cross-dataset)
ID = *all* 43 GTSRB classes
OOD = every crop from the Chinese TSRD dataset.

Within each regime we replicate the three-stage pipeline from §3: (i) train the closed-set classifier, (ii) sweep latent sizes for a vanilla autoencoder, and (iii) repeat the sweep for the Mahalanobis-enhanced variant. The zero-shot CLIP fallback is used only for the regime that works the best. At the end, we present the full system, end-to-end performance.

4.1 Same-Dataset OOD

Our first set of experiments keeps everything inside GTSRB: the OOD gate must single out the 18 *tail* classes while the 25 *head* classes are routed to the closed-set classifier. This is the most challenging configuration because tail images share colour palette, aspect ratio and iconography with the heads.

4.1.1 The Classifier

Training Figure 2 plots accuracy and cross-entropy every epoch for the ImageNet-initialised ResNet-50. Validation accuracy jumps above 98 % in six epochs and saturates at $\mathbf{99.5}\%$; the loss curves reveal no widening gap, confirming that freeze-then-unfreeze plus a weight-decay of 10^{-4} keep overfitting in check. The best model is frozen for all subsequent OOD tests.

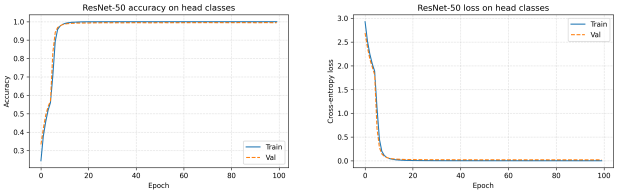


Figure 2: *Training curves for the 25-way ResNet-50. Left: accuracy. Right: loss. Solid = train, dashed = validation.*

ID-Class Test Performance Overall accuracy on the head test split is **98.06 %**; precision and recall are 97.8 % and 97.5 %. Figure 3 decomposes the result: most classes exceed 0.95 in both metrics, whereas some IDs lag behind some metric, and class 30 is the most ambiguous with the least precision and recall.

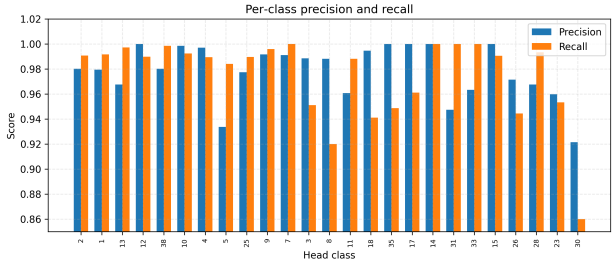


Figure 3: *Per-class precision (blue) and recall (orange). Labels are abbreviated for space; full names appear in the appendix.*

Are We Confidently Accurate? Figure 4 shows that *average* max-softmax confidence tracks true accuracy almost one-to-one across the 25 head classes: the best classes lie in the top-right, whereas class 30 occupies the lower-left, indicating that the network *knows* when a prediction is unreliable, an encouraging property for the downstream OOD gate.

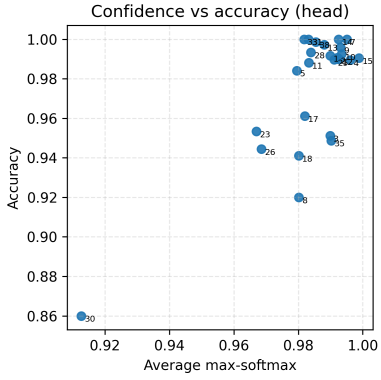
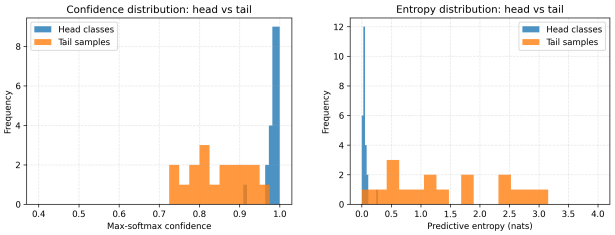


Figure 4: *Head-class accuracy versus average confidence. Points are labelled with the abbreviated class name.*

ID *vs.* OOD Distributions Without an explicit "unknown" label the network nevertheless hesitates on tail

images. The overlaid histograms in Figure 5 show a stark separation:

- **Confidence**: head predictions cluster tightly above 0.99, whereas tail samples form a broad bump peaking at ≈ 0.78.
- **Entropy**: head classes seldom exceed 0.10 nats (natural-log units), but tail samples spread out to nearly 1.5 nats.



(a) Max-softmax confidence *(b) Predictive entropy*

Figure 5: *Head (blue) versus tail (orange) uncertainty statistics.*

So, where do the tail images go? The row-normalised confusion matrix (Figure 6) shows that the head classes 38, 35, and 33 absorb a lot of the tail classes; interestingly, the model confuses almost always the (unseen) class 39 for the (seen) class 38; inspecting the classes, class 38 is a "mandatory right turn" sign, while class 39 is a "mandatory to turn left" sign. The only difference between them is the direction of the arrow, so it is normal they're confused.

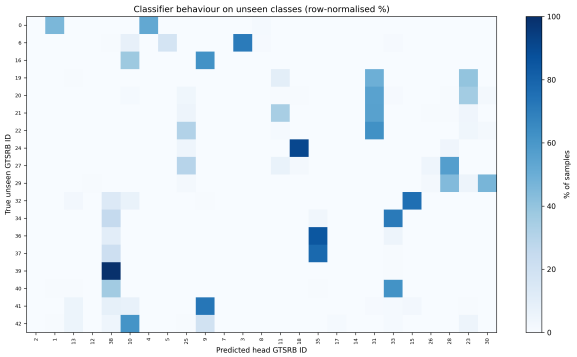
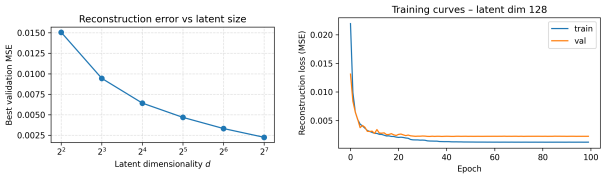


Figure 6: *Row-normalised confusion matrix: each tail class (row) expressed as a percentage distribution over the 25 head predictions (columns).*

4.1.2 Training The Autoencoder

For both the "vanilla" (plain reconstruction error) and "improved" (+ Mahalanobis distance) autoencoders, we train them the same autoencoder first. Figure 7a traces the minimum validation MSE as the latent width grows; performance improves quickly up to d=16 and then levels off, so we keep the widest model (d=128) for the final analysis. Its learning curve in Fig 7b reveals fast, stable convergence and a small but persistent gap between train and validation error, with little overfitting



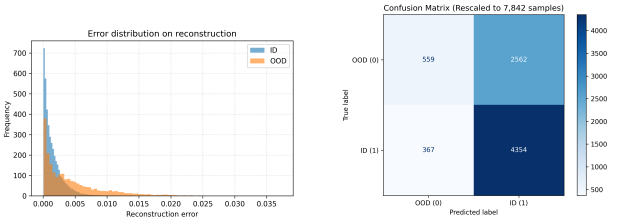
(a) *Best validation MSE versus latent dimensionality ($\log_2$ scale on d).* **(b)** *Training curves for d=128. Both losses saturate below 3×10^{-3}.*

Figure 7: *Autoencoder performance analysis.*

This autoencoder will serve as the base for the following two subsections.

4.1.3 Vanilla Autoencoder

Our first detector is the plain reconstruction-error gate. When ID-validation and tail OOD errors are overlaid (Fig 8a) the two histograms almost coincide; the 95-th percentile threshold that rejects 5 % of ID images therefore sits dangerously close to the bulk of the tail distribution. As a result, only 1 284 of the 7 170 OOD samples (18 %) are actually rejected, whereas 843 head images are mistakenly flagged, as summarised by the confusion matrix in Fig 8b. In terms of aggregate metrics the model reaches the model reaches $F_1 = 0.75$ (precision 0.63 / recall 0.92.



(a) *Reconstruction-error histogram, d=128. Blue = ID (validation); orange = tail OOD.* **(b)** *Confusion matrix after the 95-percentile ID threshold.*

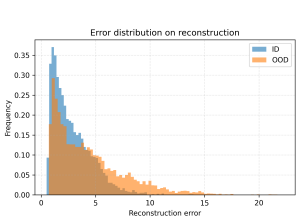
Figure 8: *OOD detection with reconstruction error.*

These findings motivate a representation-level check: many tail signs lie close to the latent manifold centre even though they are semantically "unknown". The next section therefore augments the very same autoencoder with a Mahalanobis distance in latent space, hoping to penalise samples that reconstruct well but wander too far from the ID cluster.

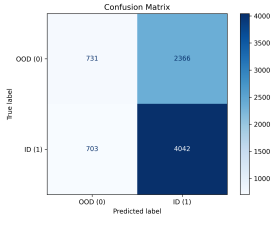
4.1.4 Mahalanobis-enhanced autoencoder

Re-utilising once again the same base autoencoder, we added the Mahalanobis distance in latent space. The ID and OOD histograms in Figure 9a still overlap, yet the

tail distribution is now markedly shifted to the right; consequently the 95-percentile ID threshold rejects a larger share of OOD samples.



(a) *Reconstruction error w. Mahalanobis filter, d=128. Blue = ID (validation), orange = tail OOD.*



(b) *Confusion matrix at the 95-percentile ID threshold.*

Figure 9: *OOD detection with Mahalanobis filter.*

Using the confusion–matrix counts we obtain $\text{precision} = 0.63,\ \text{recall} = 0.85,\ F_1 = 0.72$. Relative to the pixel-only gate the Mahalanobis test keeps precision almost unchanged yet sheds $\approx 7\%$ recall, so the resulting F_1 slips from 0.75 to 0.72.

4.1.5 Vanilla *vs.* Mahalanobis

Table 2 summarises the head–tail results. Adding the latent Mahalanobis check *removes* 409 false positives (ID precision stays put at 0.63) but *creates* 783 new false negatives, shrinking recall from 0.92 to 0.85 and pulling the F_1-score down from 0.75 to 0.72. In short, the representation-level distance trades recall for a small cut in false alarms; the pixel-only reconstruction gate remains the more balanced choice unless higher precision is the sole priority.

Table 2: *Head–tail OOD metrics, d=128.*

	Precision	Recall	F_1
Vanilla AE	0.63	0.92	0.75
+ Mahalanobis AE	0.63	0.85	0.72

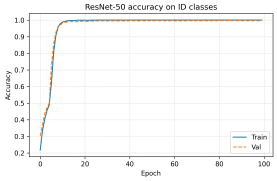
In the next section we repeat the same analysis on the *foreign-sign* scenario to see whether the gap widens when OOD samples differ not only by class identity but also by iconography and capture conditions.

4.2 Foreign-sign OOD

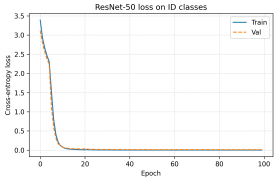
In the foreign-sign setting we broaden the ID to the *full* 43-class GTSRB taxonomy and treat every image from TSRD as OOD. The classifier therefore faces the easier closed-set task of recognising *all* German signs, while the detector must now distinguish German from Chinese iconography.

4.2.1 The Classifier

Training Figures 10a and 10b show that the ImageNet-initialised ResNet-50 converges as fast as the one in §4.1.1. Validation accuracy reaches 99.8 % after ten epochs and saturates, while cross-entropy loss below 10^{-3} with no divergence between train and validation curves, indicating that there is no overfitting.



(a) *Accuracy.*



(b) *Loss.*

Figure 10: *Foreign-sign classifier learning curves. Train and validation lines overlap almost perfectly.*

ID-Class Test Performance Figure 11 breaks accuracy down by class. Precision and recall exceed 0.95 for most categories; the handful of lower bars correspond to what we believe are visually ambiguous signs such as class 22 (“warning: bump ahead”), 42 (“no longer forbidden to pass”), or 43 (“no longer forbidden to pass trucks”). It is likely we confuse these “no longer forbidden” with their “forbidden” counterparts. The near-perfect performance across the spectrum shows that any error, once the whole system is set up, will most likely stem from the detector, not from closed-set misclassifications.

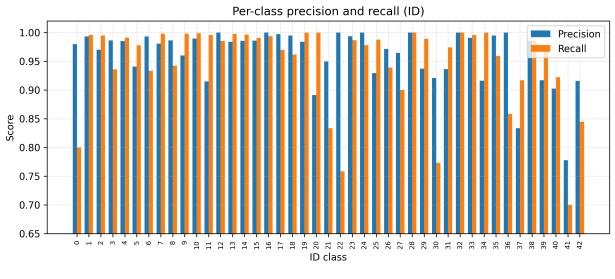


Figure 11: *Per-class precision (blue) and recall (orange) on the 43-way German test split. Most bars sit above 0.95; notable dips appear at classes 22, 30, 37, 42 and 43.*

Are We Confidently Accurate? To check whether the softmax scores remain well calibrated after expanding to 43 classes, we plot the average max-softmax confidence against true accuracy for every class (Figure 12). Most points cluster tightly along the diagonal, indicating that the model’s confidence is an honest proxy for correctness; classes with perfect accuracy ($y = 1$) also report near-unit confidence. A few outliers appear in the lower-left corner: class 22 shows high confidence and low accuracy, for example. This overall hints that

over-confidence is restricted to the foreign TSRD signs, not to any particular German category.

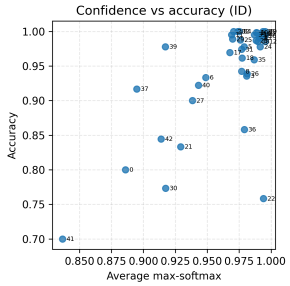
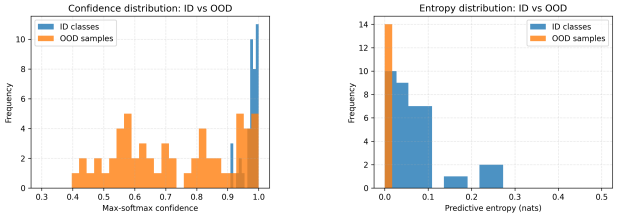


Figure 12: *Per-class average confidence versus accuracy on the 43 German test classes. Labels denote the class index.*

ID *vs.* OOD Distributions Inspection of the entropy histogram in Figure 13b reveals that *all* TSRD predictions cluster around 0 nats, lower even than the bulk of German ID samples. In other words, the classifier is *over-confident* on Chinese signs, assigning all probability mass to some German class rather than showing the hesitation we observed in the head–tail setting. This reinforces why a dedicated OOD gate is necessary despite the near-perfect closed-set accuracy.



(a) *Confidence distribution.* **(b)** *Entropy distribution.*

Figure 13: *Predictive uncertainty on ID (blue) versus TSRD OOD (orange).*

So, where do Chinese signs land? The row-normalised heat-map in Figure 14 reveals how the classifier maps Chinese to German signs. We believe, although this would require further inspection that the network leans on coarse shape and colour cues that transcend national standards to make the predictions.

Compared with the head–tail regime, the closed-set accuracy ceiling is even higher. For OOD classes, however, entropy drops near 0 nats, confirming the classifier is over-confident on Chinese signs. We now retrain the autoencoder for the vanilla and Mahalanobis detectors.

4.2.2 Training the Autoencoder

We once again train the base autoencoder for both the "vanilla" and "improved" version. The reconstruction sweep (Figure 15a) follows the same diminishing-returns pattern seen in the head–tail regime: a four-fold reduction in error from $d{=}4$ to $d{=}64$ and a much smaller

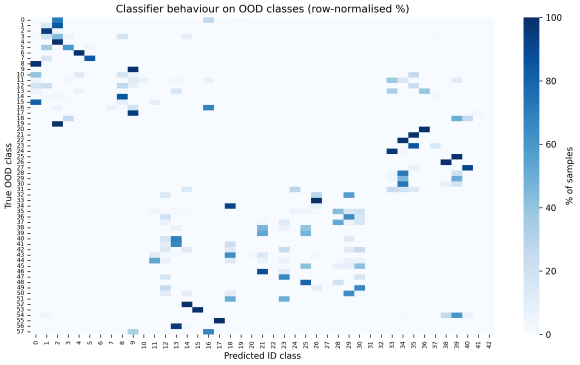
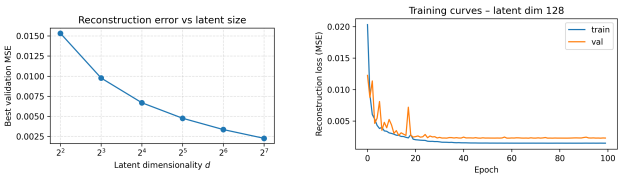


Figure 14: *Classifier behaviour on TSRD: each row sums to 100 %. Columns are German ID classes.*

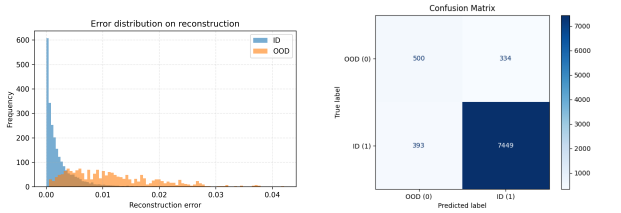
gain thereafter. With $d{=}128$ the model converges in < 30 epochs; validation loss settles around 2.5×10^{-3} without signs of overfit (Figure 15b).



(a) *Best validation MSE vs. latent width ($\log_2$-scaled d).* **(b)** *Learning curves, $d{=}128$.*

Figure 15: *Foreign-sign vanilla autoencoder. Validation loss still shrinks with larger d, yet the shoulder already flattens past $d{=}64$; we again keep $d{=}128$.*

4.2.3 Vanilla autoencoder



(a) *Error histogram, $d{=}128$.* **(b)** *Confusion matrix at the 95-percentile ID threshold.*

Figure 16: *Foreign-sign OOD detection with plain reconstruction error.*

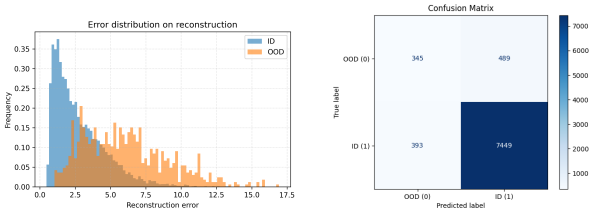
Unlike the intra-dataset case, the ID and TSRD histograms in Figure 16a now separate cleanly: Chinese signs form a long tail an order of magnitude beyond the German cluster. Applying the same unsupervised 95^{th}-percentile cut therefore rejects many more OOD images. From the confusion matrix in Figure 16b we obtain $\text{precision} = 0.96,\ \text{recall} = 0.95,\ F_1 = 0.95$.[2]

[2]Positive class = ID.

Compared with the head–tail regime both recall ($0.92 \rightarrow 0.95$) and precision ($0.64 \rightarrow 0.96$) climb substantially, lifting F_1 by twenty points (0.75→0.95). We attribute the gain to genuine visual differences between German and Chinese iconography: the autoencoder can still reproduce familiar circular or triangular silhouettes, but struggles with TSRD glyphs, colours and aspect ratios, thereby yielding markedly larger reconstruction errors. Plain reconstruction hence suffices once OOD samples drift far from the ID manifold; the next subsection tests whether the Mahalanobis filter can add value on top of this stronger baseline.

4.2.4 Mahalanobis-Enhanced Autoencoder

Replacing the scalar reconstruction test with the joint rule described in § 3.2 (MSE + Mahalanobis criteria) now shifted the OOD landscape in an *unhelpful* direction. The empirical 95-percentile threshold computed on ID-validation now sits well inside the orange bulk of foreign signs (Figure 17a); many Chinese crops still reconstruct cleanly *and* fall close to the latent mean, so the Mahalanobis guard lets them slip through. Confusion matrix counts in Figure 17b still translate into impressive figures, $\text{precision} = 0.94,\ \text{recall} = 0.95,\ F_1 = 0.94$.



(a) *Error histogram (d=128).* **(b)** *Confusion matrix at 95-percentile ID cut-off.*

Figure 17: *Foreign-sign OOD with latent Mahalanobis filter.*

4.2.5 Vanilla *vs.* Mahalanobis, Foreign-Sign

Table 3 shows the results for both detectors in the cross-dataset setting. Compared with the vanilla gate, the latent distance now trims both precision (1%) and recall (−1%), pushing the F_1 down by one point; the extra latent test therefore *hurts* in this cross-dataset scenario; although the difference is so small, we must remember the Mahalanobis variant is *more expensive*, and it clearly does not justify its cost. The mismatch suggests that foreign signs already lie far outside the pixel manifold—plain reconstruction is a strong cue, whereas the latent Gaussian assumption is too crude to model the diversity of TSRD crops.

Reaching this point, we select the vanilla autoencoder, in the foreign-sign set-up to continue building our system. On the next section, we will experiment with the zero-shot prediction.

Table 3: *Foreign-sign OOD metrics (d=128).*

Detector	Precision	Recall	F_1
Vanilla AE	0.95	0.96	0.95
+ Mahalanobis AE	0.94	0.95	0.94

4.3 Zero-shot Prediction

Finally, we assess how well a frozen CLIP back-end can *directly* label foreign signs—independently of the OOD gate. All images in the official TSRD train split are forwarded to the CLIP module described in §3.3; each sample is scored against 58 class prompts and assigned the Top-1 match.

Table 4: *Zero-shot CLIP Top-1 metrics on the complete TSRD test set.*

	Precision	Recall	F_1
CLIP Top-1	35.73%	40.88%	38.13%

With no task-specific training, CLIP retrieves the correct Chinese label for roughly four out of ten signs. While the absolute $F_1 = 0.38$ is modest compared with closed-set German accuracy, it provides a useful first guess whenever the OOD gate fires.

Note that in the full system pipeline CLIP is queried *only* for images already flagged OOD by the auto-encoder. Because that subset differs distributionally from the data evaluated here, the operational precision–recall balance may shift slightly, as we will explore in the following subsection.

4.4 Full System Performance

Finally, we evaluate the full system performance, using for the first time both ID's and OOD's test sets. At deployment time every input follows the two-stage path previously described in Eq. 1. For each image we record both the the gate decision (ID vs. OOD) and the latter class prediction (43 German or 58 Chinese labels).

Recall that a threshold on the reconstruction error is required to classify an input as either an in-distribution (ID) or out-of-distribution (OOD) sample. This threshold was determined during the training phase and corresponds to the 95^{th} percentile of the validation reconstruction error, which was 0.00828. Consequently, if an image exhibits a reconstruction error greater than 0.00828, it is classified as OOD; otherwise, it is considered an ID sample.

We will measure two metrics, both the gate performance and the end-to-end performance. The former measures binary effectiveness of the auto-encoder to separate the two domains, and we report both accuracy and F_1 (positive = ID). For the latter, a single prediction is correct only if *both* the gate and the branch classifier

get it right. We quote overall Top-1 accuracy and F_1 over the 101 labels (43 + 58).

Table 5: *End-to-end results on the held-out German (ID) and Chinese (OOD) test sets.*

Metric	Result
OOD-gate accuracy	**94.07 %**
OOD-gate F_1	**94.23 %**
System Top-1 accuracy	**88.25 %**
System F_1	**61.92 %**

Although the gate already attains a strong $F_1 \approx 94\%$, the F_1 of the full system settles at $\approx 62\%$. The gap mostly reflects two error sources: (i) the 5% of images that still cross the gate incorrectly, and (ii) residual CLIP mislabellings inside the OOD branch.

5 Discussion

When comparing results between *head–tail* and *foreign-sign* we found out the nature of the unknown classes sharply OOD-detector behaviour. When rare German signs are treated as OOD the pixel-based auto-encoder struggles, because those "tail" samples still share colour palette, shape and capture conditions with the in-distribution set. In contrast, Chinese TSRD signs break these low-level regularities; the very same architecture, trained and calibrated in exactly the same way, now flags them far more reliably. After thresholding at the 95^{th} percentile of the validation error, precision climbs from 0.63 to 0.95 while recall remains on the mid 90s, so the F_1-score jumps from 0.75 to 0.95. Table 6 shows this.

Table 6: *Side-by-side comparison of the four detectors (d=128; positive class = ID).*

Regime	Detector	Precision	Recall	F_1
Head–Tail	Vanilla AE	0.63	0.92	0.75
	+ Mahalanobis	0.63	0.85	0.72
Foreign-sign	Vanilla AE	**0.95**	**0.96**	**0.95**
	+ Mahalanobis	0.94	0.95	0.94

Adding a latent Mahalanobis distance is beneficial only when the pixel-space margin is thin, and even then whether it offers an improvement is arguable. Inside GTSRB, where reconstruction errors for head and tail images overlap, the additional representation test suppresses false alarms at the cost of a modest recall drop, trimming the F_1 by three hundredths. Once the margin widens, as with Chinese signs, the latent filter removes genuine outliers along with noise and therefore hurts both precision and recall, confirming that a single Gaussian fit is too crude for the heterogeneous TSRD encodings.

The zero-shot CLIP fallback, despite reaching only 38% Top-1 accuracy on the full TSRD test split, still proves valuable because it operates exclusively on samples the gate has already rejected.

Several positive outcomes emerge. A plain reconstruction auto-encoder, trained with no OOD data and calibrated with a single unsupervised percentile, already offers competitive open-set performance when the distributional shift is pronounced. The closed-set ResNet-50 attains $99.8\,\%$ F_1 SCORE with only standard transfer-learning hyper-parameters, and its predictive confidence remains well aligned with correctness.

Nonetheless, limitations persist. The head–tail split exposes the brittleness of pixel-level reconstruction as soon as ID and OOD overlap substantially. The latent Gaussian assumption is too restrictive for the diverse geometry of TSRD codes, leading to missed detections. CLIP, in turn, confuses fine-grained symbols because the static prompts lack contextual nuance, and the fixed 95-percentile threshold still lets sixteen per cent of all inputs cross the gate incorrectly, capping end-to-end accuracy at $88.25\,\%$ and F_1 at 61.92%.

Future work should therefore explore feature-space AEs that reconstruct deep embeddings rather than raw RGB, learn adaptive thresholds that combine reconstruction error with predictive entropy, pre-train the AE with self-supervised objectives such as SimCLR or DINO to sharpen the in-distribution manifold, automate prompt generation and ensembling for the zero-shot branch, and finally integrate a continual-learning loop that adds newly discovered classes without full retraining.

6 Conclusion

This work presented a two-stage open-set pipeline for traffic-sign recognition in which an unsupervised convolutional auto-encoder acts as an out-of-distribution gate before forwarding inliers to a fine-tuned ResNet-50 and handing outliers to a zero-shot CLIP head. Experiments reveal that the difficulty of novelty detection depends on the visual gap between known and unknown data: foreign Chinese signs are rejected with an F_1 of 0.95%, whereas rare German signs remain challenging. Latent Mahalanobis filtering offers a modest benefit only when pixel-level margins are tight and may even harm performance when the margin is already wide. CLIP, although imperfect, supplies useful best-effort labels for rejected images and improves overall macro-F_1 without additional supervision.

The study confirms that simple reconstruction models still hold their own in safety-critical scenarios where training OOD exemplars are unavailable, and that off-the-shelf zero-shot vision-language models can plug the semantic gap left by abstention. The remaining error largely stems from threshold rigidity and limited zero-shot accuracy, both tractable with richer density estimators, adaptive calibration and prompt engineering. Addressing these points should bring open-set traffic-sign recognition closer to the reliability demanded by autonomous-driving applications.

Bibliography

[1] Taylor Denouden et al. *Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance*. 2018. arXiv: 1812.02765 [cs.LG]. URL: https://arxiv.org/abs/1812.02765.

[2] Dan Hendrycks and Kevin Gimpel. *A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks*. 2018. arXiv: 1610.02136 [cs.NE]. URL: https://arxiv.org/abs/1610.02136.

[3] Sebastian Houben et al. "Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark". In: *International Joint Conference on Neural Networks*. 1288. 2013.

[4] LinLin Huang. *Chinese Traffic Sign Database*. http://faculty.bjtu.edu.cn/7139/. Beijing Jiaotong University. Supported by NSFC Grant 61271306. Beijing, China, 2012.

[5] Gabriel Ilharco et al. *OpenCLIP*. Version 0.1. If you use this software, please cite it as below. July 2021. DOI: 10.5281/zenodo.5143773. URL: https://doi.org/10.5281/zenodo.5143773.

[6] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[7] J. Stallkamp et al. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: *Neural Networks* 32 (2012). Selected Papers from IJCNN 2011, pp. 323–332. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2012.02.016. URL: https://www.sciencedirect.com/science/article/pii/S0893608012000457.

[8] Yibo Zhou. *Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection*. 2023. arXiv: 2203.02194 [cs.CV]. URL: https://arxiv.org/abs/2203.02194.