Research Square

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

Training for motion blur in compact traffic sign classifiers

Zihua Zhu

1120245536@mail.jxust.edu.cn

Jiangxi University of Science and Technology

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11657595/v1

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

Additional Declarations: No competing interests reported.

Training for motion blur in compact traffic sign classifiers

Zihua Zhu[1*]

[1*]School of Carbon Neutrality, Jiangxi University of Science and Technology, Nanchang 330013, China.

Corresponding author(s). E-mail(s): 1120245536@mail.jxust.edu.cn;

Abstract

A traffic sign classifier trained on blurred images may respond differently as the direction of motion blur changes. We compared two image augmentations and two architectural changes using a German training-validation split that kept all views of each physical sign together. Across three seeds, training with axial blur increased macro recall averaged over three specified corruptions from 71.69% to 86.68% on 12,630 German test crops. Accuracy on the original crops fell from 98.31% to 97.64%. After removing a training track with eight exact test duplicates and refitting the models, the gain was 13.91 percentage points. Channel attention and anti-aliased pooling yielded no further gain. We then evaluated the frozen models at eight blur orientations using seven samples spanning six pixels at each angle. The gain over baseline averaged 16.74 points for horizontal and vertical blur and 4.18 points for the two diagonals. The diagonal-minus-axial difference in these gains was −12.57 points in Germany and −7.46 points across 298 physical signs in Belgium (95% sign-bootstrap interval −9.52 to −5.43). On the original Belgian crops, the selected policy reduced recall by 1.53 points. Under combined downsampling and axial blur, it increased recall by 14.90 points. Axial augmentation improved mean recall at every tested orientation in both cohorts, with smaller gains for diagonal motion.

Keywords: Traffic sign recognition, Motion blur, Data augmentation, Anti-aliasing, Corruption robustness

1 Introduction

A driver approaching a speed limit sign may encounter successive views with different blur and apparent size. At a distance, the red border may occupy several pixels while the stroke that distinguishes a 30 from a 50 occupies only one. A classifier can achieve high accuracy on unmodified sign crops even if its decisions depend on that stroke. The German Traffic Sign Recognition Benchmark (GTSRB) provides repeated views of individual signs and an official test partition for studying this problem [1].

Compact sign classifiers have previously been optimized for accuracy and computational cost [2]. The Challenging Unreal and Real Environments for Traffic Sign Recognition (CURE-TSR) benchmark examined multiple acquisition and simulation conditions and found that image degradations vary in how much they impair recognition [3]. Rubio et al. compared learned and engineered features on altered GTSRB images [4]. Common-corruption benchmarks report performance by corruption to expose failure modes obscured by a single clean score [5]. The TS-1M preprint by Zhao et al. (2026) treats low clarity and cross-region transfer as distinct recognition challenges.

Two interventions address different parts of the imaging process. Exposing a classifier to blurred examples during training changes the

evidence it learns to recognize. Low-pass filtering before spatial decimation changes the features retained in the network [6]. Squeeze-and-excitation (SE) reweights feature channels using pooled context [7]. To distinguish these effects, we measured the training policy first, then compared architectural changes under that policy. We asked whether gains under simulated degradation persisted across blur directions, network backbones and countries. In all comparisons, the unit of prediction was an individual sign crop.

2 Methods

2.1 Data and preprocessing

GTSRB supplies 39,209 training images and 12,630 official test images in 43 classes. Within each training class, we used seed 20261107 to shuffle track identifiers and reserved approximately 15% of tracks for validation. The resulting partitions contained 33,359 images from 1,112 training tracks and 5,850 images from 195 validation tracks. All 43 classes were represented in both partitions. We kept the full border supplied with each image and used bilinear interpolation to resize each red, green and blue (RGB) image to 32×32 pixels. Pixel values were scaled to $[0, 1]$ and normalized with a fixed offset of 0.4 and scale of 0.25. No image was horizontally flipped, so left and right arrows kept their meaning.

A post hoc integrity check compared resized pixels across splits and inspected native RGB arrays for exact matches. We recalculated the primary endpoint after excluding matching test images. We also removed the implicated physical track from training and refit both baseline and axial models for three seeds using the original validation set, training budget and test conditions. These six fits were separate from the 27 main runs.

2.2 Training comparisons

The compact classifier used five 3×3 convolutions with 32, 48, 48, 80 and 80 channels. Each convolution was followed by batch normalization and a rectified linear unit. Three max-pooling operations reduced the spatial size to 4×4. A 128-unit fully connected layer, dropout of 0.1 and a 43-way output completed the 297,675-parameter network. Every run started from random weights and used AdamW, batch size 256, weight decay 0.01 and 20 epochs. A cosine schedule reduced the learning rate from 0.0015 to 0.000045. We retained the weights from the final epoch and paired seeds 17, 43 and 89 across conditions. Training used bfloat16 autocasting on a GeForce RTX 5070.

All four augmentation arms included affine perturbations and brightness variation. Rotation was drawn within 10° of zero, scale within 0.9–1.1, and translation within 0.09 in normalized grid coordinates. The first additional transform downsampled an image to 12, 16 or 20 pixels per side with area interpolation and then restored its size bilinearly. The second applied a horizontal or vertical box average of length 3, 5 or 7 with replicated borders. Each active transform was applied independently with probability 0.5 per image. Size, kernel length and axis were sampled per batch. When both transforms were applied, downsampling preceded blur.

In the second factorial comparison, we held the validation-selected augmentation fixed. SE followed the last convolution, with global average pooling and an $80 \rightarrow 10 \rightarrow 80$ sigmoid gate. Anti-aliased pooling replaced each stride-two max pool with a stride-one max pool, a fixed separable $[1, 2, 1]/4$ filter and stride-two decimation with reflected padding. All other layers were identical across the four configurations. The unchanged architecture reused the three matching checkpoints from the augmentation experiment. A secondary sensitivity check trained a CIFAR-style ResNet-20 [8] with 271,867 parameters under the baseline and axial policies. Its three stages had widths 16, 32 and 64, with identity shortcuts, spatial subsampling and zero-padded channels at transitions. These six additional runs were excluded from model selection.

2.3 Endpoints and model selection

We evaluated the original crop and four degraded versions. These used 12-pixel downsampling, seven-pixel axial blur, downsampling followed by axial blur, or nine-pixel diagonal blur. Axial orientation alternated by image index in filename order. The diagonal kernel averaged pixels along one diagonal. Its length and orientation both differed from the training kernels. For condition c, macro recall M_c gave each of the 43 classes equal

weight. The primary endpoint was

$$R = (M_{\text{low}} + M_{\text{axial}} + M_{\text{combined}})/3. \quad (1)$$

Each seed produced its own score. We report means, sample standard deviations and paired seed differences. A configuration qualified if its mean validation macro recall on original crops was at least the first-stage baseline minus one percentage point. Among those qualifying, the largest validation R determined the choice. We retained the checkpoint for the seed with the median validation R for the selected configuration. The architectural extension was set after the augmentation selection. Both stages have local, dated records.

2.4 External evaluation

BelgiumTSC provides a distinct collection of annotated sign crops with physical sign identifiers [9]. Its official test archive contains 2,520 images. No Belgian training images were used, and inference retained the German classifier's full 43-way output. Two comparisons used different prespecified cohorts and weightings.

The external check of the selected augmentation mapped 17 Belgian labels to German equivalents by reviewing official pictograms and example crops. Eleven classes had at least five physical signs each. Their 882 views represented 298 signs. We calculated the fraction of correctly classified views for each sign, then averaged these fractions across signs within each class. Classes received equal weight. We computed paired percentile intervals by resampling physical signs within class 5,000 times and averaging the paired differences over the three fixed seeds. These intervals describe variation among signs in the included classes, conditional on the fitted models.

The architectural check used a separate shortlist of named signs fixed before inference. Seven classes had an unambiguous code and at least one test image. They comprised road work, yield, stop, no entry, no overtaking, roundabout and priority road. Generic speed-limit and mixed-direction codes could not identify the required German labels. The traffic-signal code had no test images. We scored all 519 images from 177 signs in the seven classes, weighting images equally within each class. This subset overlaps the 11-class cohort. Its role was to compare all configurations on the same original crops.

2.5 Blur orientation at fixed length

The earlier axial and diagonal tests changed both direction and span. We specified a post hoc control before generating predictions for it. Each original German test crop and each of the 882 crops in the 11-class Belgian cohort was evaluated at all eight angles, in 22.5° steps from 0° to 157.5°. Angles run clockwise from the horizontal in image coordinates. Seven equal-weight samples lay at offsets $t = -3, \dots, 3$ along each line. Their centers spanned six pixels at every angle. We kept weights, image scale, border replication and scoring fixed and applied blur alone.

We distributed each line sample onto the pixel grid with the separable centered quadratic B-spline basis [10]. Each resulting kernel had unit mass, zero centroid and covariance $4uu^{\top} + I/4$ in pixel-squared units, where $u = (\cos\theta, \sin\theta)^{\top}$. Thus the variances along and across the line were 4.25 and 0.25 at every angle. The horizontal and vertical views of this renderer share the same slight smoothing as its oblique views. We also repeated the German control with bilinear weights, whose axial cases reproduce the original seven-sample filter. Kernel weights and independent numerical checks accompany the predictions.

For each seed and policy, we contrasted mean recall on the two diagonals (45°, 135°) with the mean on the two axes (0°, 90°). Subtracting the baseline contrast from the axial-training contrast gives the change in the augmentation advantage. German sensitivity analyses used the previously identified duplicate-free test subset and the previously fitted checkpoints trained without the overlapping track. Belgian endpoints used the same sign-balanced weights as the 11-class comparison. To estimate paired intervals, we resampled physical signs within each class 5,000 times, retaining all angles, policies and seeds for each sampled sign.

2.6 Use of artificial intelligence

Zihua Zhu independently drafted the full text of the present manuscript and used OpenAI Codex to edit the English for grammar and clarity. During the research phase, Codex assisted in structuring the experimental workflows and preparing the experimental protocols. It implemented

Table 1 Selection on the German validation split and performance on the official test set. Values are percentages. Test R is the mean across three seeds with sample standard deviation. The eligibility floor was 96.96% original validation macro recall. The shared axial row is the control for the three architectural changes

Training and architecture	Parameters	Val. original	Val. R	Eligible	Test R
Geometry and brightness only	297,675	97.96	73.29	Yes	71.69 ± 0.95
Add downsampling	297,675	97.10	83.22	Yes	82.61 ± 0.39
Add axial blur	297,675	97.69	87.44	Yes	$\mathbf{86.68} \pm 0.61$
Add both augmentations	297,675	96.11	88.22	No	87.55 ± 0.46
Axial blur with SE	299,365	97.14	86.53	Yes	86.31 ± 0.67
Axial blur with anti-aliased pooling	297,675	96.35	85.81	No	84.41 ± 0.25
Axial blur with both modules	299,365	96.32	85.45	No	84.38 ± 0.41

and ran the training and evaluation scripts, analyzed the saved predictions, prepared the tables, and plotted the figures with Matplotlib. Verification scripts recalculated the reported scores and external intervals from the saved predictions and checked the checkpoint files. Zhu personally verified the experimental plans, reported results, and model checkpoints.

3 Results

3.1 The gain depended on the degradation

The combined training policy achieved the highest validation R but fell below the original-crop floor. Axial augmentation alone met the floor and was selected (Table 1). It improved test R by 15.71, 13.99 and 15.26 percentage points for seeds 17, 43 and 89. The mean gain was 14.99 points. The largest change occurred under combined downsampling and axial blur. Macro recall rose from 58.82% to 80.20% (Table 2). On original images, macro recall fell from 97.13% to 96.24%, and overall accuracy fell from 98.31% to 97.64%.

The seven-pixel axial test was close to the training distribution. The unfamiliar diagonal kernel yielded 47.82% macro recall under axial training, compared with 45.39% at baseline. The seeds differed noticeably for this contrast (Fig. 1). On the 2,743 original crops with a native short side of at most 32 pixels, mean overall accuracy fell from 97.10% to 96.60%. For the 9,887 larger original crops, it fell from 98.65% to 97.93%. Native small size alone did not reproduce the large benefit measured after artificial degradation.

The integrity check found eight test images that exactly reproduced native RGB frames from a single stop-sign training track. Removing those eight test crops left 12,622 images and a mean gain of 14.99 points. Removing all 30 frames from that training track left 33,329 training crops. After refitting, baseline and axial test scores were 72.42% and 86.33%, a gain of 13.91 points on the full official test set. No resized crop matched exactly between the German fit and validation sets or between the German fit set and the 2,520 Belgian test images.

3.2 Architectural changes under axial training

The unchanged network had higher validation R than SE, anti-aliased pooling or their combination. On the German test set, adding SE changed mean R by -0.37 points. Anti-aliased pooling changed it by -2.27 points, and the combination by -2.30 points. Every paired seed difference for these three comparisons was negative. The reference ResNet-20 reproduced the training effect. Its mean test R rose from 61.87% to 85.34% with axial augmentation, while original macro recall moved from 96.44% to 95.86%. The backbone choice affected the size of the gain.

The selected compact model (seed 43) correctly classified 3,072 jointly degraded images that the matched baseline had misclassified. It misclassified 312 images that the baseline had classified correctly. Both networks missed 1,638 images. Among the selected model's errors, 107 images with a 50 km/h sign were classified as 30 km/h and 100 images of 30 km/h signs were classified as 50 km/h. These are frame counts. A single physical sign can contribute several similar frames.

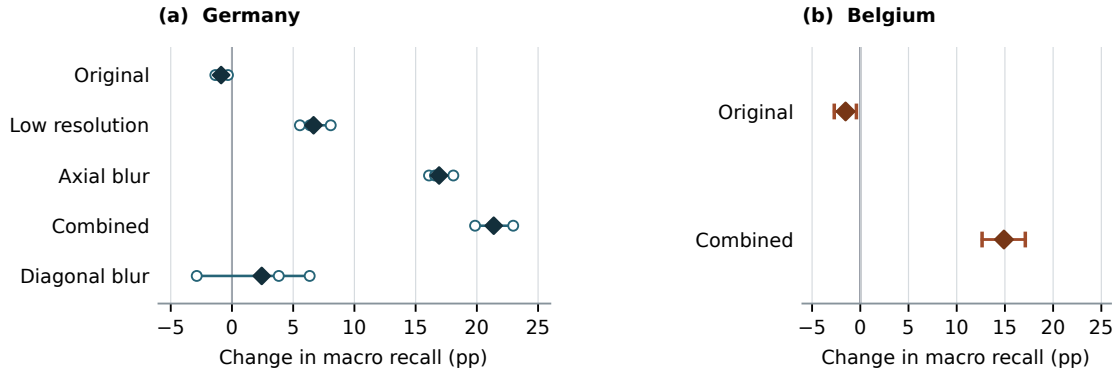

Fig. 1 Effect of axial augmentation relative to the matched compact baseline. In Germany (a), open circles are the three paired seed differences and diamonds are their means. In Belgium (b), diamonds are differences in sign-balanced macro recall across 11 classes and bars are 95% paired sign-bootstrap intervals. The vertical line marks zero change

Table 2 Mean German test macro recall for each condition and external original-crop macro recall for the seven-class Belgian subset. Results are percentages over three seeds. Joint denotes 12-pixel downsampling followed by seven-pixel axial blur

Configuration	Original	12 px	Axial	Joint	Diagonal	Belgium
Compact baseline	97.13	82.25	74.01	58.82	45.39	98.56
Downsampling	95.89	91.32	80.17	76.34	43.05	98.59
Axial blur	96.24	88.91	90.92	80.20	47.82	98.61
Both augmentations	94.76	90.26	89.21	83.17	46.47	97.96
Axial with SE	96.22	88.69	90.22	80.02	47.75	98.95
Axial with anti-aliased pooling	95.25	86.67	89.27	77.30	43.79	98.79
Axial with both modules	94.89	86.54	89.18	77.41	42.42	98.41
ResNet-20 baseline	96.44	77.11	60.91	47.59	20.37	97.58
ResNet-20 with axial blur	95.86	87.72	90.89	77.42	29.68	96.91

3.3 Performance on original and degraded Belgian crops

In the 11-class external cohort, the selected training policy reduced sign-balanced macro recall on original crops from 95.07% to 93.55%. The paired change was −1.53 points (95% interval −2.70 to −0.40). When the same Belgian images were downsampled and axially blurred, the score increased from 71.32% to 86.22%. The change was +14.90 points (95% interval +12.63 to +17.11). Figure 1 shows why a single external average would hide an important difference between the two conditions.

All seven compact configurations scored at least 97.96% on the narrower seven-class comparison of original crops (Table 2). Here the axial and baseline models differed by only +0.05 points in macro recall. The seven-class selection excludes several warning categories present in the broader analysis, and its weighting differs. It cannot estimate performance over all 62 Belgian labels or distinguish close model rankings reliably.

3.4 Axial training retained a smaller advantage on diagonals

At a fixed blur span, German macro recall for the classifier trained with axial blur averaged 90.15% on the two axes and 79.43% on the two diagonals (Table 3). The diagonal-minus-axis changes for the three seeds were −11.41, −10.23 and −10.53 points. Baseline recall rose from 73.41% to 75.25%. The augmentation advantage therefore fell from

Table 3 Fixed-span blur by angle for the frozen compact classifiers. Entries are three-seed mean macro recall (%) for Germany and sign-balanced macro recall (%) for the 11-class Belgian cohort. Gains are paired percentage-point differences calculated before rounding. Each crop appears at every angle

	Germany (12,630 crops)			Belgium (298 signs)		
Angle	Baseline	Axial training	Gain	Baseline	Axial training	Gain
0°	75.98	90.96	14.98	76.98	86.92	9.94
22.5°	77.77	87.09	9.33	72.29	82.14	9.85
45°	78.67	81.18	2.50	75.26	78.84	3.58
67.5°	73.44	83.28	9.84	82.64	88.15	5.51
90°	70.84	89.35	18.51	82.61	93.21	10.60
112.5°	68.16	83.38	15.22	90.20	92.35	2.15
135°	71.84	77.69	5.85	88.71	90.76	2.05
157.5°	78.92	86.03	7.11	85.49	88.39	2.89

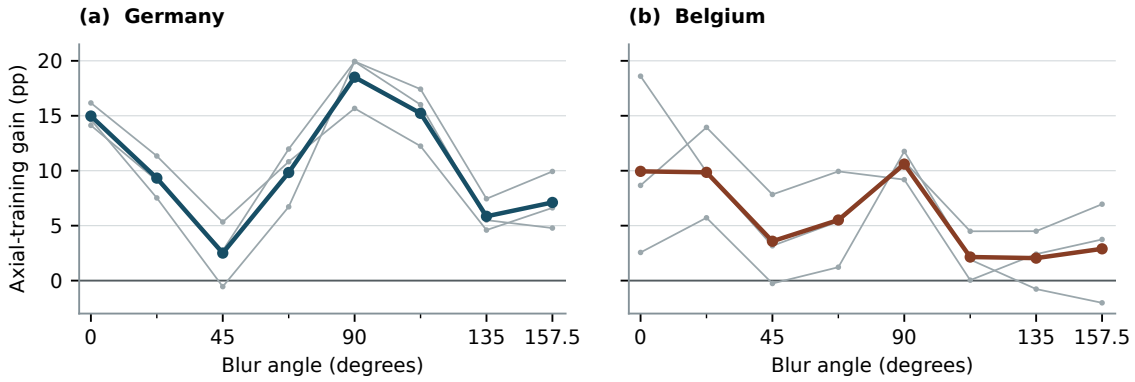

Fig. 2 Paired gain in macro recall from axial training at the same six-pixel span. Thin lines show the three paired seed differences and heavy lines with circles show their mean. Segments connect the eight measured angles. German scores weight classes equally, and Belgian scores also weight physical signs equally within class

16.74 points on the axes to 4.18 on the diagonals, a paired difference of −12.57 points before rounding. The mean gain stayed positive at every angle, although seed 89 lost 0.55 points at 45° (Fig. 2).

The change in advantage was −12.57 points after excluding the eight exact duplicate test crops, −13.75 with the checkpoints fitted without the overlapping training track, and −13.00 with bilinear rasterization. Alternating the bilinear horizontal and vertical predictions in original filename order reproduced all 75,780 archived axial predictions from the six German checkpoints.

In Belgium, recall for the classifier trained with axial blur fell from 90.06% on the axes to 84.80% on the diagonals, a change of −5.27 points (95% sign-bootstrap interval −6.52 to −3.99). The baseline gained 2.19 points. The training advantage was 10.27 points on the axes and 2.81 points on the diagonals. Its paired change was −7.46 points (95% interval −9.52 to −5.43). The two diagonal orientations differed substantially in the Belgian cohort, which is why both are reported.

4 Discussion

Exposure to axial blur helped two backbones classify crops subjected to a related loss of detail. Adding SE or spatial low-pass filtering provided no further primary-endpoint gain for the compact model. This experiment assessed one SE placement and one pooling filter. The results do not establish what other placements, kernels or training budgets would achieve. In particular, the pooling module acts on intermediate features, whereas the augmentation changes the input conditional on the sign label.

The costs appear in settings where the augmentation no longer matches the task. Accuracy fell in both size strata of the original German crops, and sign-balanced recall fell on original Belgian crops in the broader cohort. The fixed-span control shows that orientation itself affects the benefit of axial training under synthetic motion. Its gain on diagonal inputs narrowed in both countries, with different scores for the two diagonals. The original nine-site diagonal spans $8\sqrt{2}$ pixels between sample centers and uses grid-aligned taps. Its 47.82% recall and the fixed-span 45° result of 81.18% differ in both span and sampling. We therefore cannot attribute the difference in recall to blur length alone. Matching the first two spatial moments controls blur scale across angles while leaving higher-order pixel-grid effects possible. The recurring 30/50 confusion is consistent with loss of fine numeral detail, although these predictions do not identify which pixels determined a decision.

The score R weights three synthetic conditions equally. Their frequencies on real roads are unknown. The Belgian transfer test changes traffic sign design and capture conditions together, and its degraded views remain synthetic. The German final release gives no track IDs. An exact RGB-and-dimensions comparison found only eight matches to the older sorted competition archive, so no image-independent confidence interval is reported for the German results. Exact matching detects shared pixels but cannot exclude near-duplicate views. Seed standard deviations describe training variation. The Belgian intervals describe uncertainty across the observed signs and do not account for model retraining or variation between countries. No sign detector, road video sequence or vehicle decision system was evaluated.

5 Conclusion

With a fixed minimum for validation macro recall on original crops, axial training improved German corruption-averaged recall by 14.99 points across three seeds. The comparable residual network also improved, while the two structural additions reduced that score in the compact classifier. At a common six-pixel span, axial training improved mean recall at all eight angles in both cohorts, but its advantage on diagonals was 12.57 points smaller than on the axes in Germany and 7.46 points smaller in Belgium. Evaluating only horizontal and vertical motion would overstate the benefit across blur orientations.

Data and code availability

The benchmark authors distribute the final training images, test images and labels through the GTSRB archive [1]. Belgian crops and pictograms are available from the BelgiumTS maintainers [9]. The results and code supporting this article are provided in Online Resource 1.

Online Resource 1. Reproducibility archive containing source code, data locations and checksums, split assignments, per-run metrics, trained weights, rasterized kernels, per-angle predictions, the scores reported by angle, and scripts for recalculating the reported scores and external intervals.

Statements and Declarations

Funding. No funding was received for this study.

Competing interests. The author declares no competing interests.

Author contributions. Zihua Zhu independently drafted the manuscript and verified the experimental plans, reported results, and model checkpoints.

References

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

[2] Aghdam, H.H., Heravi, E.J., Puig, D. A practical and highly optimized convolutional neural network for classifying traffic signs in real-time. Int. J. Comput. Vis. **122**, 246–269 (2017). https://doi.org/10.1007/s11263-016-0955-9

[3] Temel, D., Kwon, G., Prabhushankar, M., AlRegib, G. CURE-TSR: Challenging unreal and real environments for traffic sign recognition. NIPS

Workshop on Machine Learning for Intelligent Transportation Systems (2017). https://doi.org/10.48550/arXiv.1712.02463

[4] Rubio, A., Demoor, G., Chalmé, S., Sutton-Charani, N., Magnier, B. Sensitivity analysis of traffic sign recognition to image alteration and training data size. Information **15**, 621 (2024). https://doi.org/10.3390/info15100621

[5] Hendrycks, D., Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. ICLR (2019). https://doi.org/10.48550/arXiv.1903.12261

[6] Zhang, R. Making convolutional networks shift-invariant again. Proc. ICML, PMLR **97**, 7324–7334 (2019). https://proceedings.mlr.press/v97/zhang19a.html

[7] Hu, J., Shen, L., Sun, G. Squeeze-and-excitation networks. Proc. CVPR, 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[8] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. Proc. CVPR, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[9] Timofte, R., Zimmermann, K., Van Gool, L. Multi-view traffic sign detection, recognition, and 3D localisation. Mach. Vis. Appl. **25**, 633–647 (2014). https://doi.org/10.1007/s00138-011-0391-3

[10] Unser, M. Splines: A perfect fit for signal and image processing. IEEE Signal Process. Mag. **16**(6), 22–38 (1999). https://doi.org/10.1109/79.799930

Supplementary Files

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

- ESM1.zip