SAME OBJECT, DIFFERENT VERDICT: CLASS-SPECIFIC LOCATION PRIORS IN SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Moving an object can change a segmentation model's verdict even when its pixels are unchanged. We study these class-specific location preferences through object placements, controlled training, and label-free prediction maps. On 400 reserved ADE20K sources, 65–67% of objects that reach IoU ≥ 0.5 somewhere fall below it at another placement in the same background. The effect's direction agrees with the class's training layout for 69–72% of objects with nonzero vertical contrast. Correlated training induces comparable placement penalties in zero-padded and exactly circularly equivariant CNNs: joint translation equivariance does not prevent learning object–background relationships. Architecture changes the preference's persistence under corruption and its visibility in average prediction maps. Finally, translation augmentation reduces whole-scene shift penalties by 55% and 38% in DeepLabV3+ and SegFormer, while largely preserving their class-specific placement profiles. A better scene-shift score therefore does not certify object-placement robustness. Evaluating spatial reliability requires both, alongside the class-directed effects that pooled averages conceal.

1 INTRODUCTION

Semantic segmentation assigns meaning to pixels, but those assignments can depend on where the pixels occur. This is often useful: ceilings tend to be above floors, and cars usually appear near the ground. The same regularities become a liability when a recognisable object occurs elsewhere. In Figure 1, moving an object within a fixed scene changes its segmentation despite leaving its appearance unchanged.

A familiar response is to improve translation robustness. However, two different interventions are easily conflated. Translating a complete image moves the object and its surroundings together. Relocating only the object changes its relationship to the surrounding scene. Even an exactly translation-equivariant predictor may respond differently to the latter. Moreover, averaging placement effects across classes can conceal them: an upward preference for trees and a downward preference for cars can cancel. A model can therefore look robust under a pooled shift metric while retaining systematic class-specific preferences.

We connect four experiments around this distinction: identical-object placements on ADE20K, controlled manipulation of class–location associations during training, label-free maps of spatial prediction structure, and translation-augmentation finetuning. Together, they link observed placement effects to training layouts, test the role of image boundaries, and establish whether improving a scene-shift benchmark also changes the class-specific placement profiles.

The evidence supports three conclusions. **Placement effects are large, class-directed, and related to training layouts.** Approximately two thirds of test objects recognised at one placement cross below IoU 0.5 elsewhere. **Exact joint-translation equivariance permits learned placement dependence.** Correlated training induces nearly the same clean placement penalty in circular and zero-padded CNNs, although their behaviour under corruption differs. **Better scene-shift scores leave class-specific placement profiles largely intact.** Translation augmentation improves the former in both deployed model families while preserving the latter's structure. We use *location prior* operationally for a learned class-specific spatial preference, without assuming an explicit probabilistic prior or a unique internal mechanism.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

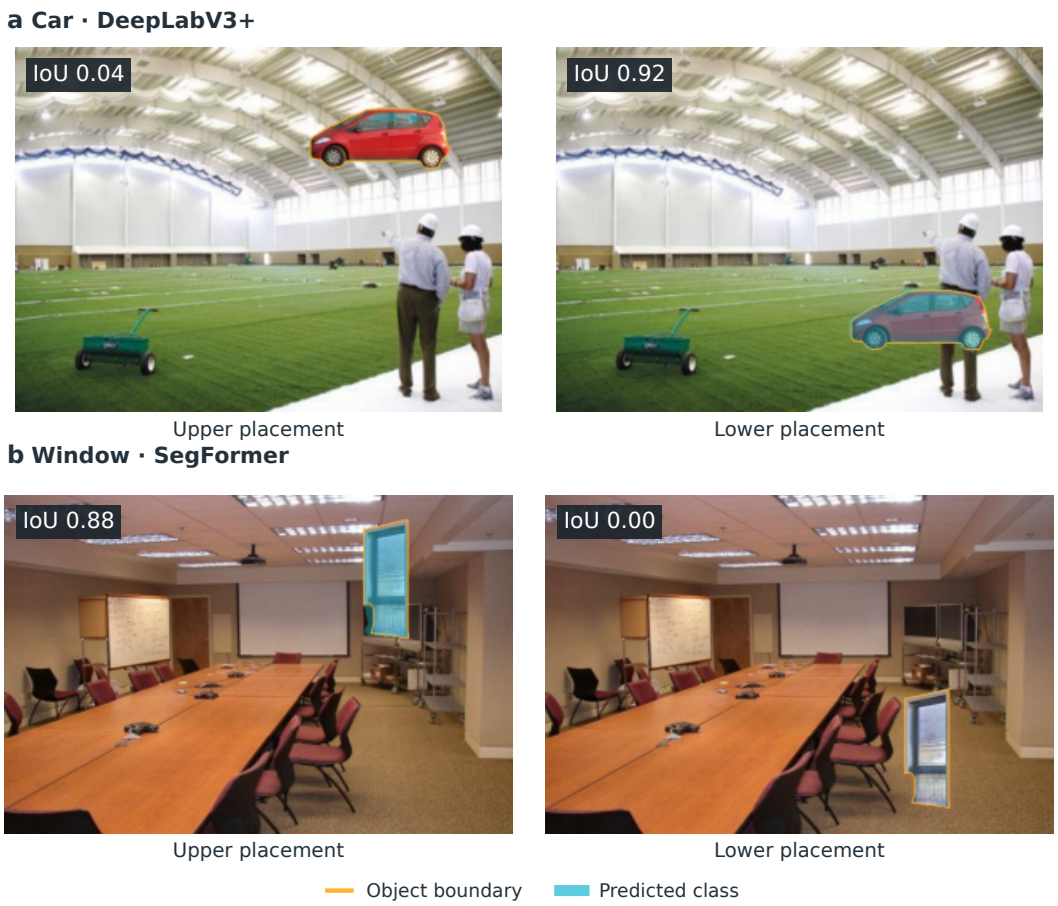


Figure 1: **Identical pixels, different segmentation.** Within each pair, the same object is moved vertically on the same background, without scaling or blending. DeepLabV3+ nearly loses the car at the upper position; SegFormer loses the window at the lower position. Orange outlines mark the pasted object and cyan marks predictions of its class. These examples were selected to illustrate large effects; Table 1 gives their frequency on the reserved test set. All five placements and both models appear in Figure 7.

2 Related work

Position information and boundaries. CNN features can encode absolute position despite weight sharing (Islam et al., 2020; Kayhan & van Gemert, 2020). Padding shapes this information and can introduce spatial blind spots (Alsallakh et al., 2021); its effects extend to semantic representations and downstream predictions (Islam et al., 2021b). Position can also survive global pooling and contribute to segmentation (Islam et al., 2021a). Lin et al. (2023) isolate position-information patterns from image features. Thus, the question is not simply whether a network knows where a feature occurs. We measure which classes benefit from which placements, test whether those preferences follow the training distribution, and distinguish object relocation from joint translation.

Context and spatial robustness. Scene context supports object recognition (Torralba, 2003; Oliva & Torralba, 2007). Object transplanting already exposes position-dependent detection failures (Rosenfeld et al., 2018), and object-removal interventions reveal contextual dependencies in classification and segmentation (Shetty et al., 2019). Small image transformations expose failures of expected invariance (Azulay & Weiss, 2019; Engstrom et al., 2019). Anti-aliasing improves CNN shift robustness (Zhang, 2019), and architectural changes can make vision transformers equivariant to circular shifts (Rojas-Gomez et al., 2024). These guarantees concern specified transformations of the entire

input. We connect placement failures to class-specific training layouts and show how they coexist with exact equivariance and improved scene-shift scores.

Spatial structure in transformers. Position-related feature artefacts can be reduced by feature denoising (Yang et al., 2024) or, in DINO models, by adapting positional encodings (Pawlowsky et al., 2026). Training-only coordinate supervision provides another way to shape spatial organisation (Liu et al., 2026). These findings motivate separating architectural position signals from spatial relationships learned through data. Our deployed transformer, SegFormer, uses padded convolutions without an explicit positional embedding (Xie et al., 2021); its observed location dependence cannot be attributed to an absolute embedding table. Our output-level maps complement feature analyses, and their interpretation is anchored by training interventions and a symmetry null.

3 Separating object placement from scene translation

Models and sources. We evaluate public ADE20K checkpoints of DeepLabV3+ with a ResNet-50 backbone (Chen et al., 2018) and SegFormer-B2 (Xie et al., 2021). From the ADE20K validation set (Zhou et al., 2017), a fixed rule selects 600 images with one eligible connected component each, spanning 105 classes. Components occupy 0.5–20% of the image, contain at least 150 pixels, and do not touch the frame. The manifest separates 200 development and 400 test sources. Headline recognition and alignment results use the reserved test sources; descriptive class profiles use all 600. Appendix A specifies selection, inference, and scoring.

Object placement. We paste the selected component, with its RGB pixels unchanged, onto a distinct background image containing no annotation of its class. The background is fixed across five placements: centre and the four quarter/three-quarter positions. Placement is clamped to keep the component fully visible. For source i and position u, let $J_i(u)$ be the selected class's intersection-over-union (IoU), including false positives elsewhere in the canvas. Define its vertical placement contrast as

$$d_i = \tfrac{1}{2}\big[J_i(\mathrm{TL}) + J_i(\mathrm{TR}) - J_i(\mathrm{BL}) - J_i(\mathrm{BR})\big]. \tag{1}$$

Positive d_i means better segmentation in the upper half. An object is *recognisable somewhere* if $\max_u J_i(u) \geq 0.5$; a *verdict flip* occurs if it also has $\min_u J_i(u) < 0.5$.

Scene translation. We separately shift the original image and its annotation together by the four diagonal displacements $(\pm 0.15W, \pm 0.15H)$. Exposed image pixels are zero-filled and exposed labels ignored. Shifted IoU is measured on visible, valid support; the unshifted reference uses the original support. Here, the selected-class score can include multiple instances. These shifts preserve local scene relationships where visible, but introduce truncation and borders. They are therefore a complementary benchmark, not a matched causal isolation of context.

Training layout and uncertainty. For each class c, we estimate a normalised spatial density G_c from all 20,210 ADE20K training annotations and extract a top-minus-bottom preference g_c using four windows centred on the placement positions. The aligned effect is $a_i = \mathrm{sign}(g_{c_i})d_i$: positive values favour the training layout. We report both its mean over all test objects and sign agreement among nonzero contrasts. Confidence intervals use 5,000 bootstrap replicates; class-level claims additionally resample classes and permute one preference per class, preserving all within-class dependencies, with 9,999 permutations. Controlled-study intervals instead cluster repeated compositions by foreground source. No analysis treats pixels as replicates.

4 Placement effects follow class-specific training layouts

Large changes are common among recognisable objects. On the 400 test sources, moving an object below the recognition threshold occurs for 55 of 85 recognisable objects in DeepLabV3+ and 55 of 82 in SegFormer (Table 1). Across all recognisable objects, the mean best-to-worst IoU ranges are 0.476 and 0.509. These rates condition on demonstrated recognition: the model segments the object successfully at one location, then loses that recognition when only its placement changes.

Training layouts predict the direction. Among test objects with nonzero d_i, the training-layout sign agrees with the performance sign for 126/176 objects (71.6%) in DeepLabV3+ and 110/159 (69.2%) in SegFormer. Their conditional mean aligned effects are $+0.103$ and $+0.112$ IoU, with

Table 1: **Recognition depends on placement.** The denominator is the number of test objects reaching IoU ≥ 0.5 at any of the five placements. Most test objects never reach this threshold. A zero IoU is no intersection with the target, not necessarily no predicted pixels.

Model	Recognisable / 400	Flip below 0.5	Reach zero IoU
DeepLabV3+	85	55 / 85 (64.7%)	35 / 85 (41.2%)
SegFormer	82	55 / 82 (67.1%)	28 / 82 (34.1%)

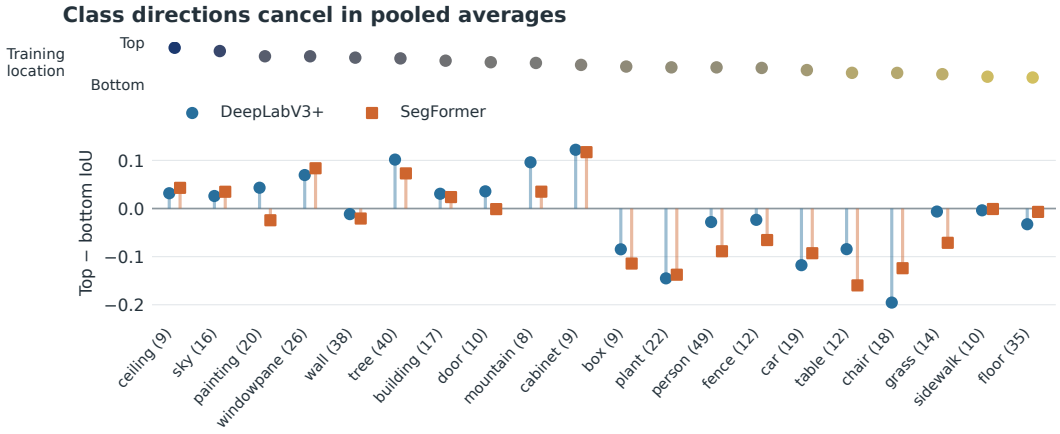


Figure 2: **The direction depends on the class.** Mean top-minus-bottom placement IoU for every class with at least eight of the 600 sources; counts are in parentheses. Classes are ordered by their mean vertical position in ADE20K training annotations, shown above. Upper-image classes tend to benefit from upper placements and lower-image classes from lower placements, with exceptions. Opposing directions can cancel in a pooled average. Formal alignment tests use the separate 400-source test subset and all eligible classes, not only those displayed.

class-cluster 95% intervals $[0.058, 0.157]$ and $[0.061, 0.153]$. The remaining 224 and 241 objects have zero vertical contrast; retaining them gives mean aligned effects of $+0.045$ in both models. Class-preserving permutation tests give $p = 0.0001$ and $p = 0.0010$. Estimating the training density after simulated crop and flip augmentation yields the same direction of evidence (Appendix B). This establishes alignment with training statistics, not a causal mechanism; the next section supplies a training intervention.

Pooled shifts obscure the structure. Across all 600 sources, the average off-centre minus centre IoU is only $+0.003$ for DeepLabV3+ and -0.011 for SegFormer, despite the class profiles in Figure 2. Whole-scene translations yield larger pooled losses, 0.051 and 0.034, but these depend strongly on retained support: among the 80 scenes retaining all selected-class pixels at all four shifts, mean changes are -0.009 and $+0.001$, with intervals spanning zero (Figure 8). Thus, neither a pooled placement score nor an unstratified scene-shift score captures the class-directed effect measured by d_i.

5 Training and Architecture Shape Placement Preferences

Intervening on the class–location association. We construct 128×128 scenes from 12 COCO foreground classes (Lin et al., 2014) and natural-image backgrounds. Each class is assigned a quadrant. In *uniform* training, positions are sampled across the canvas; in *correlated* training, 80% of draws target the assigned quadrant and 20% use the uniform rule. Foregrounds, backgrounds, scales, and training recipes are shared across conditions, with source images disjoint across training, development, and test. Matched CNNs use zero or circular padding. Both contain only stride-one convolutions and spatially shared pointwise operations, making the circular model exactly equivariant to integer circular translations. Matched transformers use learned absolute positional embeddings or no positional embeddings.

Test compositions reuse each foreground–background pair under three placement rules: uniform, *favoured* (the training mixture), and *reversed* (the same mixture, targeting the opposite quadrant). Reversal changes the targeted 80% branch, not every composition. Each of the eight checkpoints is evaluated on 2,400 compositions from 240 held-out foreground instances and 40 backgrounds. The fits share one random seed. Intervals cluster by foreground instance, conditional on these checkpoints and the background pool; they do not estimate training-seed variation. Appendix C gives the architectures, recipes, and checkpoint lengths.

Exact joint-translation equivariance permits a learned placement penalty. After correlated training, mean IoU falls from 0.490 in the favoured layout to 0.330 in the reversed layout for the zero-padded CNN, and from 0.488 to 0.337 for the circular CNN. Their uniform-trained controls show little separation. Subtracting that control contrast defines the excess layout penalty:

$$L = \left(\bar{J}_{\text{fav}} - \bar{J}_{\text{rev}}\right)_{\text{correlated}} - \left(\bar{J}_{\text{fav}} - \bar{J}_{\text{rev}}\right)_{\text{uniform}}. \tag{2}$$

We obtain $L = 0.156$ $[0.140, 0.173]$ and 0.155 $[0.140, 0.172]$, respectively. Changing the training layout creates a comparable effect in both CNNs. Moving an object while fixing the background is not a joint translation: the circular model can still use object–background relationships correlated with the assigned layout. These results establish that such a route is available, without identifying its internal features.

The transformer control distinguishes positional access from stronger symmetry. With absolute embeddings, correlated training gives favoured/reversed IoUs of 0.152/0.062 and $L = 0.093$ $[0.080, 0.106]$. Without positional embeddings, the corresponding scores are 0.051/0.051 and $L = 0.002$ $[-0.003, 0.006]$. The latter model is equivariant to arbitrary permutations of its 4×4 patches, removing their relative arrangement as well as an absolute coordinate table. That restriction is stronger than the circular CNN's translation equivariance. Both transformer variants have much lower accuracy than the CNNs, so this comparison neither ranks architectural robustness nor establishes an accuracy-preserving remedy.

Corruption and probe maps separate the effects. We replace a fraction q of conjugate Fourier phase pairs, sharing the replacement across colour channels and retaining the original amplitudes before clipping. At $q = 0.25$ and 0.5, L is 0.150 and 0.079 for zero padding, versus 0.085 and 0.025 for circular padding (Figure 3b). The correlated CNN's favoured-to-reversed loss relative to favoured-layout IoU increases from 33% to 47% for zero padding and decreases from 31% to 9% for circular padding. Absolute penalties decline in both as segmentation deteriorates. At $q = 1$, IoU is near zero, making relative accuracy penalties uninformative.

After conditioning probabilities on foreground classes, the mean assigned-quadrant mass is 53% for the zero-padded correlated model on fully phase-randomised scenes, versus 24% for its uniform-trained control. The corresponding transformer values are 33% versus 24% with absolute embeddings. Circular CNNs and transformers without positional embeddings remain near 25%. A preference's visibility therefore depends on both architecture and probe distribution. Section 6 develops this diagnostic and compares phase, constant, white-noise, and spectrally filtered probes.

6 Label-free maps expose class-location structure

The controlled experiment links prediction-map structure to an imposed training layout. We now develop this as a diagnostic: vary the input distribution, average the model's responses, and test which spatial structures persist. We then apply it to deployed segmenters, where the training layout is natural and the internal mechanism is not controlled.

For class c, input law Q, and output grid Ω, define the mean response, its prevalence, and its normalised spatial shape:

$$B_c(u) = \mathbb{E}_{X \sim Q} p(c \mid X, u), \qquad \pi_c = \frac{1}{|\Omega|} \sum_u B_c(u), \qquad s_c(u) = \frac{B_c(u)}{\sum_v B_c(v)}. \tag{3}$$

These maps require neither probe annotations nor an auxiliary predictor. Normalisation separates spatial shape from class prevalence, but can amplify very weak responses. For the deployed models, we therefore retain classes with $\pi_c \geq 10^{-4}$ and report the coverage. We also subtract the mean shape

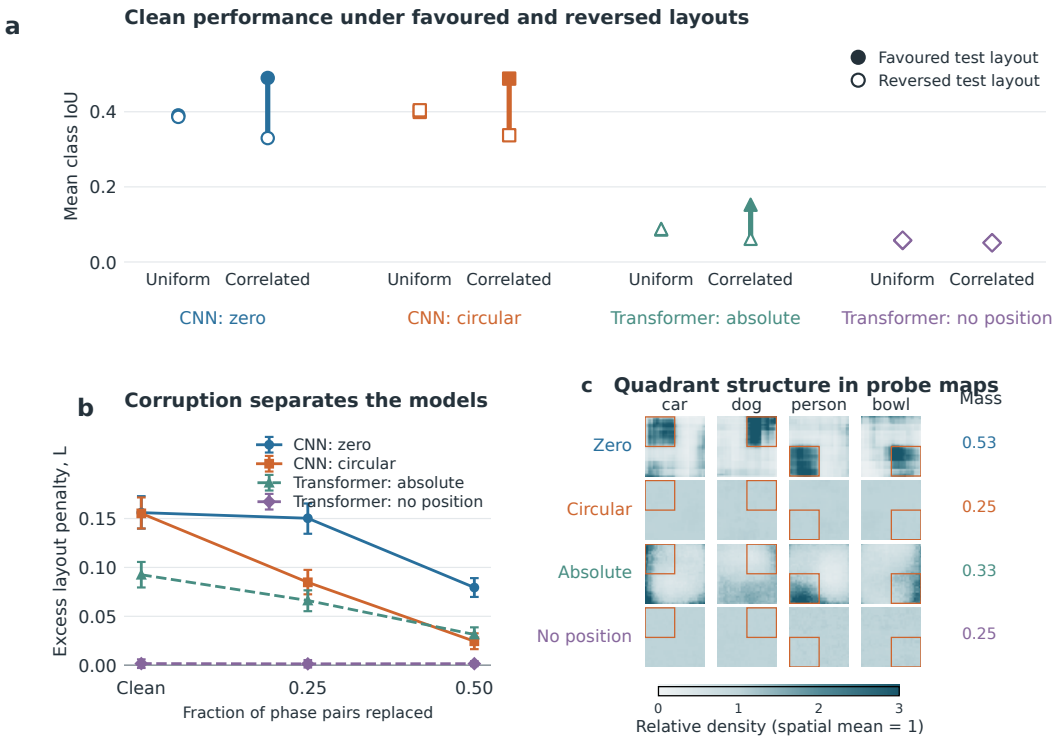


Figure 3: **Training layout and architecture jointly determine the preference.** (a) Correlated training separates favoured- and reversed-layout IoU in both CNNs and the absolute-position transformer. The transformer without positional embeddings has little separation and low accuracy. Horizontal labels indicate training layout. (b) The excess layout penalty L subtracts each uniform-trained control's contrast. Bars are 95% bootstrap intervals over 240 foreground instances. (c) Foreground-conditional maps on fully phase-randomised scenes, for correlated-trained models. Outlines mark assigned quadrants. Four classes illustrate all four quadrants; the right column gives mean assigned-quadrant mass over all 12 (uniform reference: 0.25). Maps show spatial shape, not absolute foreground confidence.

across retained classes, $r_c = s_c - \bar{s}$, to distinguish class-specific structure from a shared centre or border pattern.

The controlled models have an explicit background class that dominates many probe responses. For them, we first condition each pixel's probabilities on the 12 foreground labels, $\widetilde{p}(c \mid X, u) = p(c \mid X, u) / \sum_{k=1}^{12} p(k \mid X, u)$, then average and spatially normalise as above. This measures how foreground classes compete across positions, even when their total probability is small. Conditioning each input before averaging is essential: conditioning the average raw map would instead give greater weight to inputs with stronger foreground responses.

For example, the zero-padded correlated model's mean raw probability per foreground class is only 3.68×10^{-4} under full phase replacement and 2.09×10^{-6} under constant input. Its structured conditional maps describe relative preferences within these weak responses, not confident foreground detections.

Two ways to vary the evidence. Phase corruption starts with a scene and changes the arrangement encoded by its Fourier phases while retaining its amplitude spectrum. We replace a fraction q of conjugate phase pairs, sharing the random phase and replacement decisions across RGB channels. The original amplitudes and self-conjugate phases, including the mean (DC) component, are retained before clipping and quantisation. Figure 4a shows the resulting progression. Even at $q = 1$, spectral and colour information remains; this is a controlled change in evidence, not its complete removal.

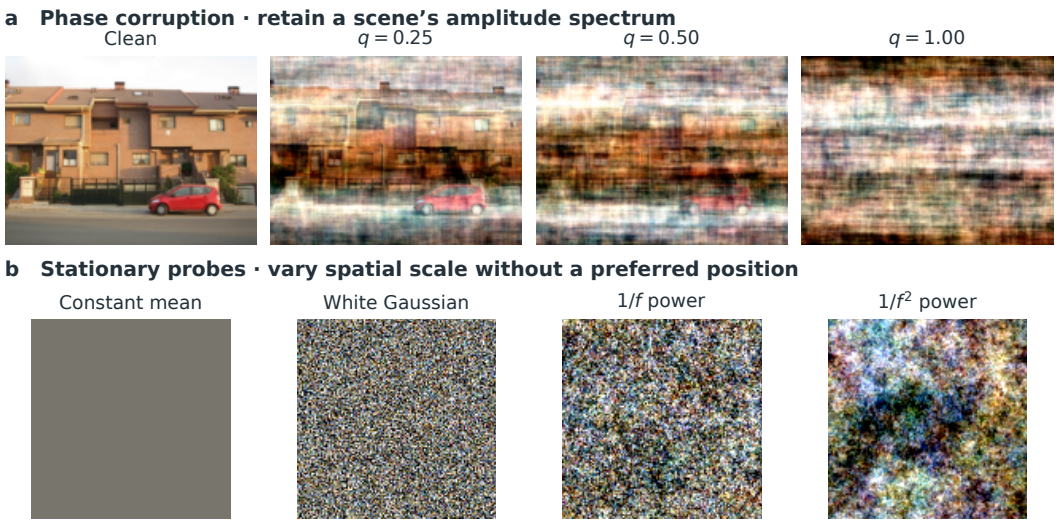


Figure 4: **Probe design separates scene structure, spatial scale, and position.** (a) One ADE20K scene under increasing phase replacement. The amplitude spectrum and self-conjugate phases are retained before RGB clipping; complete phase replacement still leaves spectral information. (b) Inputs from the controlled study's probe generators, using its training colour statistics: a constant field, white Gaussian noise, and spectrally filtered noise with power proportional to $1/f$ or $1/f^2$ before RGB conversion. Every bottom-row input law is stationary under circular translation. They can nevertheless elicit different class-location maps: mean assigned-quadrant mass across the 12 foreground-conditional class maps is 53% under full phase replacement and 23% under $1/f$ noise for the correlated-trained zero-padded CNN, versus 33% and 38% for the absolute-position transformer (uniform reference: 25%). A probe's failure to reveal a preference does not establish its absence.

Stationary noise provides a complementary intervention (Figure 4b). It contains no source scene and has the same distribution after any circular translation. White noise excites all spatial frequencies; filtering it with Fourier gain $f^{-\alpha/2}$, for radial frequency $f > 0$ and $\alpha \in \{1, 2\}$, produces the $1/f$ and $1/f^2$ power laws used here. Increasing α shifts energy towards coarser structure without introducing a preferred position. We use training-image colour moments to scale and mix the channels, then round and clip to RGB. Pointwise conversion preserves stationarity, although it can change the spectrum. The constant mean-colour field supplies a limiting comparison with no spatial variation. These probes vary texture scale and scene-specific spectral information separately from positional stationarity.

A symmetry null for the expected map. Let a finite group $\mathcal{G}$ act on inputs and spatially permute outputs. For predictor F, write $B = \mathbb{E}_Q F(X)$ for the full mean output array and $\Pi_{\mathcal{G}} B = |\mathcal{G}|^{-1} \sum_g gB$ for its group average.

Proposition 1 (Expected-map symmetry)**.** *If Q is invariant under $\mathcal{G}$ and F is integrable, then*

$$\|B - \Pi_{\mathcal{G}} B\| \leq \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{E}_Q \|F(gX) - gF(X)\|. \tag{4}$$

Thus, equivariance makes the expected map invariant; if the spatial action is transitive, each class map is constant.

The bound follows from input-law invariance, Jensen's inequality, and averaging over the group; Appendix D.2 gives the proof and subgroup cases. It supplies a null for stationary probes, not a semantic interpretation of every departure from flatness: boundaries, stride, decoding, and nonstationary inputs can all contribute. Our phase-corrupted scenes require the symmetry restrictions stated there. Alignment with class layouts, together with the controlled intervention, carries the semantic interpretation.

The input law determines which preferences become visible. Across the 12 controlled classes, the correlated-trained zero-padded CNN assigns mean foreground-conditional mass 0.533 to the

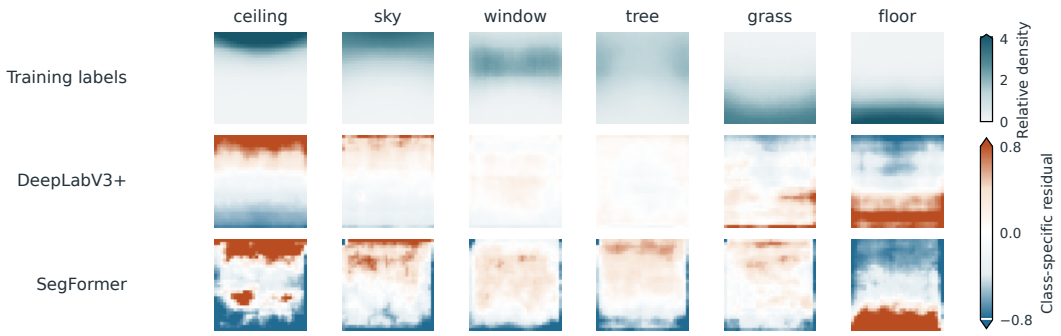


Figure 5: **Class-location structure survives severe phase corruption.** Training-label densities (top) and class-specific residual prediction maps on fully phase-randomised scenes (bottom). Six classes illustrate upper- and lower-image structure; all pass the response gate in both models. The same colour scale is used for both model rows, after normalising each map to spatial mean one and removing the shared class-average map. Training densities use a separate nonnegative scale. Quantitative alignment includes all retained classes: 150 for DeepLabV3+ and 48 for SegFormer. These are average responses to corrupted scenes, not predictions of object masks in those scenes.

training quadrants under full phase replacement and 0.561 under constant input, but only 0.233 under $1/f$ noise. The absolute-position transformer gives 0.327, 0.401, and 0.380, respectively; its uniform-trained control gives 0.243, 0.255, and 0.247. Circular CNNs and transformers without positional embeddings remain near the uniform reference of 0.25 under these probes (Table 4). These are spatial shapes, not recognition confidence. The circular CNN's flat stationary-probe maps coexist with its clean placement penalty in Figure 3: joint-translation symmetry constrains the former without removing the latter. A family of probes is therefore more informative than treating any single noise image distribution as a universal readout of learned location dependence.

Deployed maps align with natural training layouts. For each corruption level, we average 256 draws cycling through 32 ADE20K validation images and pool outputs to a 32×32 grid. With all replaceable phase pairs randomised, residual-map and training-density top-minus-bottom signs agree for 92/150 classes in DeepLabV3+ (61.3%) and 36/48 in SegFormer (75.0%). Permuting class identities gives $p = 0.0001$ for each. Figure 5 shows representative map shapes; Appendix D reports prevalence gating and the full corruption sweep. Coverage differs substantially between the models, so the percentages are not a controlled comparison of their prior strength. The result is a practical diagnostic: class-specific spatial structure is measurable in outputs even when accurate segmentation is largely lost.

The distinction from recognition is measurable. In the separate object-placement experiment on all 600 sources, full phase replacement reduces mean selected-class IoU from 0.134 to 0.003 in DeepLabV3+ and from 0.131 to 0.003 in SegFormer. Mean maximum softmax probability over the output canvas nevertheless remains 0.455 and 0.651. Neither scalar describes where particular classes receive probability. The maps reveal that spatial organisation, while the training-layout comparison tests whether it is class-specific rather than merely a shared border response.

7 Better scene-shift robustness leaves placement profiles intact

If scene translation and object placement expose different dependencies, improving one should not be assumed to solve the other. We test this directly by finetuning each deployed checkpoint for 1,000 updates, comparing the standard training pipeline with the same pipeline plus random image-and-label translations of up to 15% in each axis. Each arm uses three paired seeds, giving 12 fits. Learning rates and all other settings are matched within each model family (Appendix E).

Translation augmentation reduces the average scene-shift penalty from 0.0517 to 0.0230 in DeepLabV3+ and from 0.0364 to 0.0227 in SegFormer, improvements of 55.5% and 37.5% (Figure 6a). Every paired seed improves. Clean validation mIoU changes from 42.23 ± 0.61 to 41.82 ± 0.49

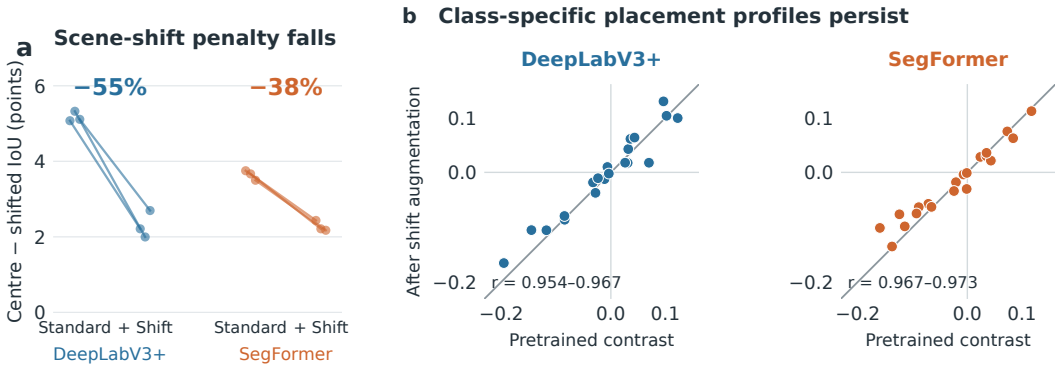


Figure 6: **A better shift score leaves the class-specific placement pattern.** (a) Translation augmentation reduces the mean scene-shift penalty by 55% for DeepLabV3+ and 38% for SegFormer. Each line joins paired training seeds; penalties average the 600 source scenes. (b) Per-class vertical placement contrasts after augmentation remain close to their pretrained values. Points average three seeds and include every class with at least eight sources; the diagonal marks equality. The displayed correlation ranges are across individual seeds. Together, the panels show why a scene-shift score cannot stand in for an object-placement audit.

and from 44.97 ± 0.08 to 44.82 ± 0.25 (mean $\pm$ sample standard deviation, percentage points). The gain therefore accompanies small observed clean-accuracy decreases.

Class-specific object-placement profiles nevertheless persist. After translation augmentation, their correlations with the pretrained profiles are 0.954–0.967 for DeepLabV3+ and 0.967–0.973 for SegFormer (Figure 6b). The root-mean-square profile magnitudes remain 90–96% and 77–86% of their pretrained values, respectively. Across the reserved test objects, the mean training-layout-aligned contrast remains positive in every augmented fit, at 0.038–0.039 and 0.033–0.037 IoU. This is a measured separation between the two robustness targets, not a claim that translation augmentation can never alter location dependence.

8 Discussion

The results distinguish three properties that are often treated together: equivariance to joint translation, robustness to object placement, and the spatial structure of class predictions. Training layouts affect placement performance even in an exactly circularly equivariant CNN. Boundary and positional information change how those preferences appear in probe maps. The transformer control shows that removing patch arrangement imposes a stronger symmetry, accompanied here by low accuracy. Translation augmentation improves a scene-shift benchmark while preserving most of the class-specific placement structure. Each measurement answers a different question.

The study has defined limits. Pasting creates unnatural scenes and changes local context along with object position; it is a controlled stress test, not an estimate of ordinary validation error. The deployed-model evidence covers two checkpoints, and the controlled training comparison uses one training seed. Phase randomisation retains spectral information and can move inputs far from the training distribution. Label-free maps reveal responses under a chosen input law, rather than a unique causal decomposition of a network.

The probe family also suggests a useful experimental direction. Spectral filtering changes texture scale while preserving translation stationarity; phase corruption instead retains a particular scene's amplitudes. Comparing the resulting response maps can help formulate hypotheses about which visual statistics activate a spatial preference. Such hypotheses can then be tested by training interventions and object placements, connecting an inexpensive diagnostic to functional behaviour.

For evaluation, the implication is concrete: report class-directed placement contrasts alongside whole-scene shift scores, retain their denominators, and check whether opposite class preferences cancel. A segmenter's tolerance to moving a scene does not tell us whether it will recognise the same object elsewhere within that scene.

REPRODUCIBILITY STATEMENT

The current supplementary material contains a separate PDF of Appendices A–E, including experimental protocols, model and training specifications, detailed result tables, uncertainty calculations, and the symmetry proof. Figure, table, and equation numbering continues from the main paper.

We will release the complete reproducibility package with the final paper upon acceptance. It will include per-example measurements; raw and foreground-conditional probability maps; source-selection and composition manifests with dataset identifiers, annotations, and hashes; all eight trained controlled-model checkpoints; controlled training and evaluation code; scripts for statistical analysis and figure generation; and editable figure and manuscript sources. Public dataset and deployed-model identifiers will accompany the package. The same materials can also be shared with reviewers and area chairs through an anonymous link during the rebuttal period. This full artifact release is separate from the present supplementary submission.

ETHICS STATEMENT

This study uses public research datasets. Spatial shortcuts can make predictions depend on scene framing, motivating evaluation beyond average accuracy. The experiments do not establish reliability for safety-critical deployment or fairness across human groups.

STATEMENT ON THE USE OF AI

This work involved substantive AI assistance with writing, retrieval and discovery, research ideation and execution, and drafting sections. The authors established the research question, mathematical formulation, experimental design, figure concepts, and initial manuscript, and retain scientific and editorial responsibility. AI tools provided methodological feedback and simulated peer review, checked algebra and assumptions, suggested corrections and notation, and assessed interpretations and limitations. Identified deficiencies sometimes motivated further experiments or checks.

AI tools supported literature discovery, novelty assessment, citation verification, and drafting and substantial revision of manuscript and supplementary text, including the abstract and captions. Coding assistance, primarily Codex, contributed to implementation, debugging, analysis and validation scripts, plotting and layout code, reformatting numerical records, and organising and anonymising the supplement.

Photographs and annotations came from existing datasets. Inputs used the stated compositing, translation, Fourier-phase, and noise-generation procedures; prediction overlays came from the evaluated models. Figure assistance operated through code, preserving experimental image content and model outputs apart from the described transformations and overlays.

Validation included correctness and numerical tests on controlled cases, checking values and figures against outcome records, and examining cited work and bibliographic details against primary sources. The authors take responsibility for the submitted text, claims, mathematics, code, data provenance, figures, artifacts, and verification statements.

REFERENCES

Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad—CNNs can develop blind spots. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02178.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019. URL https://www.jmlr.org/papers/v20/19-519.html.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1802–1811, 2019. URL https://proceedings.mlr.press/v97/engstrom19a.html.

Md Amirul Islam, Sen Jia, and Neil D. B. Bruce. How much position information do convolutional neural networks encode? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJeB36NKvB.

Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G. Derpanis, and Neil D. B. Bruce. Global pooling, more than meets the eye: Position information is encoded channel-wise in CNNs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 793–801, 2021a.

Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G. Derpanis, and Neil D. B. Bruce. Position, padding and predictions: A deeper look at position information in CNNs, 2021b. URL https://arxiv.org/abs/2101.12322. arXiv:2101.12322.

Osman Semih Kayhan and Jan C. van Gemert. On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

Chieh Hubert Lin, Hung-Yu Tseng, Hsin-Ying Lee, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unveiling the mask of position-information pattern through the mist of image features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21204–21222, 2023. URL https://proceedings.mlr.press/v202/lin23k.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Cong Liu, Xiaofang Li, and Simon X. Yang. Active spatial guidance: Eliminating injected positional mechanisms in vision transformers, 2026. URL https://arxiv.org/abs/2607.00580. arXiv:2607.00580.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007. doi: 10.1016/j.tics.2007.09.009.

Moritz Pawlowsky, Antonis Vamvakeros, Alexander Weiss, Anja Bielefeld, Samuel J. Cooper, and Ronan Docherty. What DINO saw: ALiBi positional encoding reduces positional bias in vision transformers, 2026. URL https://arxiv.org/abs/2603.16840. arXiv:2603.16840.

Renan A. Rojas-Gomez, Teck-Yian Lim, Minh N. Do, and Raymond A. Yeh. Making vision transformers truly shift-equivariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5568–5577, 2024.

Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room, 2018. URL https://arxiv.org/abs/1808.03305. arXiv:1808.03305.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk—quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8218–8226, 2019.

Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. doi: 10.1023/A:1023052124951.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiawei Yang, Katie Z. Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q. Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision*, 2024.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334, 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.