The Shape of Events: Edge-Based Inductive Biases via Cross-Domain Distillation

Soshun Kihara[1*] **Shunsuke Yasuki**[2*] **Masato Taki**[3,4*]

[1]Independent Researcher [2]The University of Tokyo, Japan

[3]Graduate School of Artificial Intelligence and Science, Rikkyo University, Japan [4]RIKEN, Japan

project page: https://github.com/snskysk/event2rgb-distillation

Abstract

Convolutional neural networks trained on ImageNet are known to exhibit a strong preference for local high-frequency texture, an inductive bias that translates into fragile robustness against distribution shifts in real-world environments. Event cameras, in contrast, record only changes in scene brightness and are therefore well suited to capturing contour information; however, due to the absence of diagnostic benchmarks in the event domain, the inductive bias that event-camera data instills in vision models has remained underexplored. In this work, we use knowledge distillation from the event domain to the RGB domain so as to exploit the rich evaluation toolkit available in the RGB domain and systematically dissect this inductive bias. Our experiments show that distillation from the event domain induces, in the RGB domain, color invariance, shape bias, and robustness to high-frequency noise. We identify the underlying mechanism as the model suppressing its dependence on high-frequency texture while acquiring a stronger dependence on edge-based object shape. This hypothesis is supported by changes in how color and spatial information are processed at the early layers, together with a spectral trade-off in which robustness to the absence of high-frequency components coexists with vulnerability to contamination of the relied-upon frequency bands and to disruption of geometric structure. We further show that this inductive bias differs from existing robustification methods and that it functions as a useful prior for diverse downstream tasks in which shape and contour information contribute alongside other cues. The code and experimental configurations for reproducing our experiments are provided alongside this paper as supplementary material.

1 Introduction

Unlike the human visual system, convolutional neural networks (CNNs) are widely reported to exhibit a texture bias, preferring local high-frequency texture over global object shape [1]. This inductive bias is an underlying cause of model fragility under real-world environmental changes such as illumination variation, noise, and domain shift. In this paper, we present a new approach to this problem from the perspective of the event-based sensor (event camera) domain.

Event cameras possess properties that conventional cameras lack, including high temporal resolution and high dynamic range [2]. Moreover, their hardware design records only per-pixel changes in brightness: uniform static regions emit no events, whereas spatial brightness gradients (including textured regions) generate events only when motion converts them into temporal changes. Edges and contours are therefore represented in the event stream more reliably than textures, whose presence is motion-dependent, and we expect models trained on event data to acquire internal representations grounded in macroscopic geometric structure and edges.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

*All authors contributed equally.

However, research on the internal representations of models trained on event data is still insufficient. In the RGB domain, benchmarks for dissecting and diagnosing models are well established, including evaluations against diverse corruptions such as ImageNet-C [3] and shape–texture cue-conflict images [1]. In the event-vision community, by contrast, evaluation metrics and diagnostic datasets that allow detailed analysis of model behavior are lacking. As a result, the representations that event data induces in vision models remain underexplored.

To overcome this evaluation barrier and dissect the representations specific to event data, we propose an approach based on the simple framework of knowledge distillation [4]. Specifically, we use a model trained on the large-scale event dataset Neuromorphic-ImageNet (N-ImageNet) [5] as a teacher, and transfer (distill) its knowledge into a student model in the RGB domain (ImageNet [6]). This cross-domain distillation has two implications. First, by transplanting the features and properties of event data into the RGB domain, we can dissect their underlying mechanism using the rich evaluation toolkit of the RGB domain. Second, by distilling event-derived constraints into an RGB model, where texture would otherwise dominate as a shortcut, we can suppress shortcut learning in a way that has not been studied before. Note that, because N-ImageNet is generated by recapturing ImageNet on a monitor, our experiments isolate the brightness-gradient (edge) selectivity property of event data; HDR and motion parallax are outside the experimental scope (Section 7).

Our experiments reveal that models distilled with knowledge from event data (hereafter, Event-distilled models) acquire a set of robustness properties (color invariance, shape bias, and resistance to high-frequency noise) compared to baseline models. Furthermore, by analyzing the high spatial correlation of input-layer filters and the model's behavior under digital artifacts (robustness to JPEG compression, in which unstructured texture is removed, and vulnerability to perturbations that destroy geometric structure such as edge continuity), we examine the mechanism behind this robustness. That is, the Event-distilled model learns an edge-based shape feature, a property distinct from prior robust models and shape-biased models.

The main contributions of this work are summarized as follows:

- Event-driven knowledge distillation: We propose an analytical use of cross-domain distillation from event- to RGB-trained models, deliberately adopting the standard knowledge-distillation formulation [4]. Our contribution lies not in a new distillation objective but in the cross-domain setup and in the systematic characterization of the resulting inductive bias through RGB-domain diagnostics.
- A characteristic set of robustness properties: Through comprehensive evaluation, we observe that the Event-distilled model exhibits a set of phenomena, including color invariance, a shift toward geometric and structural cues over texture, and improved accuracy under high-frequency noise and under adversarial perturbations with $0 < \epsilon \leq 1/255$.
- Mechanism elucidation and downstream applications: We identify that the source of these robustness properties is a shift in the internal representation, in which the model suppresses its dependence on high-frequency texture and acquires a stronger dependence on the edge-based shape feature. We further show that this property functions as a useful inductive bias for downstream tasks, suggesting new directions for event data in computer vision.

2 Related Work

2.1 Event-based Vision and Neuromorphic Datasets

Event cameras asynchronously capture brightness changes [2] and are robust to illumination change and motion blur. The recent emergence of large-scale datasets such as N-ImageNet [5, 7–18] has driven the proposal of recognition models such as DiST [5], including a wide range of subsequent approaches [19–32]; following N-ImageNet, many of them convert the event stream into a grid representation (an event image) so that existing architectures such as CNNs can be reused. Applications to other computer-vision tasks such as object detection [33–39] and semantic segmentation [40–44] are also expanding.

To compensate for the scarcity of labeled event data, distillation from a pretrained RGB model into an event model has been proposed [45–48]. To our knowledge, no work sends knowledge in the opposite direction, from an event model into an RGB model, in order to analyze the representational

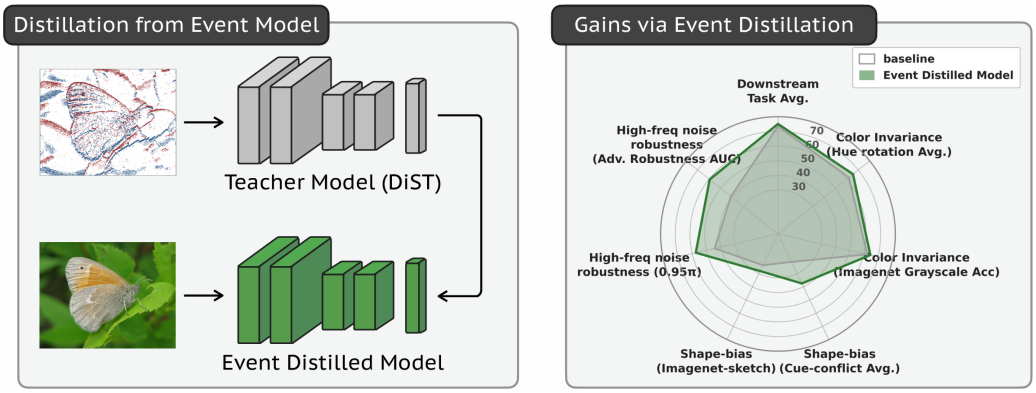


Figure 1: Method overview and summary of results. *Left*: Cross-domain knowledge distillation from a frozen DiST event teacher (trained on N-ImageNet) into an RGB ResNet-34 student (trained on ImageNet). N-ImageNet and ImageNet are paired at the instance level, enabling per-sample KL between teacher and student logits. *Right*: Summary of robustness gains of ED_Student ($\alpha = 0.2$) over the RGB baseline across the three axes studied in this paper (color invariance, shape bias, and high-frequency noise robustness) and downstream task transfer.

properties that event data induces. Hybrid sensors such as DAVIS [49] read out frames and events from the same pixel array; as natively paired data become available at scale, our dissection could be repeated on them.

2.2 Inductive Biases in CNNs: Texture vs. Shape

It has been widely pointed out that standard CNNs trained on RGB images exhibit a texture bias, preferring local high-frequency texture over the overall shape of objects [1, 50–57]. In response, shape-biased models have been trained on style-transferred images such as Stylized-ImageNet [1], which destroy texture and force the model to learn shape. Adversarial training and noise-augmentation training are also known to reduce texture reliance as a side effect. The Event-distilled model likewise exhibits a strong shape bias, but we show in Sections 6.3 and 6.4 that it is grounded in an edge-based shape feature distinct from those of Adversarial Training, Noise Training, and Stylized-ImageNet co-training (hereafter, the SIN model) [1].

2.3 Knowledge Distillation for Model Analysis

Knowledge distillation is a widely used technique that trains a student model using the predictive distribution (soft labels) of a teacher model [4, 58–62]. More recently, cross-modality distillation that transfers representations between different modalities (e.g., RGB and depth, RGB and audio) has also been investigated [47, 48, 63–66]. Our goal, however, is neither compression nor accuracy on a new task. Because diagnostic metrics are underdeveloped in the event domain, we transfer knowledge from an N-ImageNet teacher into an RGB student that has access to abundant diagnostics, repurposing distillation as an analytical tool for dissecting the inductive bias of a heterogeneous domain.

3 Method: Distilling Event-Driven Inductive Biases

Given the limited means available for directly analyzing the visual features (inductive biases) on which event-domain models rely, we propose a framework that repurposes cross-domain distillation as an analytical tool. Figure 1 provides an overview. Concretely, we adopt the DiST ResNet-34 [5] trained on N-ImageNet as the teacher model, and a ResNet-34 trained on ImageNet [6] as the student model.

Knowledge distillation (KD) preliminaries. Knowledge distillation [4] trains a student network to match the temperature-softened softmax outputs of a teacher network. In our setup the teacher is trained on event-domain data (N-ImageNet via DiST [5]), while the student learns RGB classification with distillation guidance from the teacher's event-domain predictions on paired samples.

3.1 Model Design and Training Strategy

To analyze the inductive bias with a classical CNN structure while ensuring fair comparison with prior work (DiST [5]), we adopt ResNet-34 [67] as our principal subject of analysis. We verify the architectural generality of our findings in Section 6.2, using VGG [68], ConvNeXt [69], and others. We choose ResNet-34 rather than the more common ResNet-50 for three reasons: the DiST teacher is itself a ResNet-34, which gives a same-architecture teacher–student pair; its moderate size (about 21M parameters) keeps the teacher variants and the multiple students within our compute budget; and the phenomena replicate on a ResNet-50 student (Section 6.2), the backbone used in the original cue-conflict study [1], so this choice does not restrict our conclusions.

As in standard knowledge distillation [4], the student model is trained to minimize a linear combination of the cross-entropy loss ($\mathcal{L}_{CE}$) with the ground-truth labels (hard label) and the Kullback–Leibler divergence ($\mathcal{L}_{KL}$) with the teacher model's predictive distribution (soft label). The overall loss $\mathcal{L}$ in this work is defined using a weight parameter $\alpha \in [0, 1]$ as follows.

$$\mathcal{L} = (1-\alpha)\,\mathcal{L}_{CE}\big(y,\, p_s(x_{\text{strong}})\big) + \alpha\, T^2\, \mathcal{L}_{KL}\big(p_t(x_{\text{event}}),\, p_s(x_{\text{weak}})\big) \tag{1}$$

Here, y is the ground-truth label, x_{strong} and x_{weak} denote the same RGB instance passed through a strong and a weak augmentation pipeline respectively, x_{event} is the paired event sample, and p_s and p_t are the (softmax) outputs of the student and teacher models. $\mathcal{L}_{KL}$ uses temperature-softened logits ($T = 3.0$) with the standard T^2 multiplier [4]. This dual-branch construction follows DeiT [61] and Patient Teacher [62]; see Appendix A.4.5 for the augmentation pipelines and cross-modal view alignment. To isolate the influence of the event-derived inductive bias on the RGB model, in addition to a baseline model (a ResNet-34 trained on ImageNet RGB images, i.e. $\alpha = 0$), we design the following two Event-distilled variants by varying the soft-label dependence α.

- ED_SoftLabelOnly ($\alpha = 1.0$): A student model trained using only the teacher's soft labels (soft:hard = 10:0). By completely eliminating the intervention of label information from the RGB domain, this model is expected to inherit the predictive distribution (inductive bias) of the event teacher in its purest form. However, because the DiST ResNet-34 teacher itself attains only 48.43% on N-ImageNet, the standard accuracy on clean ImageNet drops considerably from the baseline (73.9%→53.5%).
- ED_Student ($\alpha = 0.2$): A model trained with a mixture of soft and hard labels (soft:hard = 2:8). To prevent large clean-accuracy drops from confounding our analysis, we adopt the empirically standard $\alpha = 0.2$ weighting that balances regularization and knowledge transfer, preserving the event-derived inductive bias while alleviating degradation of basic image-classification capability.

The notation soft:hard refers only to the α weighting in Eq. 1 (α for the KL term and $1-\alpha$ for the CE term), before the standard T^2 scaling of the KL loss. With $T = 3$ the effective coefficient of the KL term is therefore $\alpha\, T^2 = 1.8$ for ED_Student, but we keep the 2:8 notation for the α weighting itself, following the convention of prior knowledge-distillation work.

Since our objective is mechanism analysis rather than SOTA accuracy, we do not exhaustively tune the soft/hard label mixing ratio. Detailed training settings including data augmentation are described in Appendix A.

4 Phenomena: Robustness Properties of Event-Distilled Models

We apply the Event-distilled models (ED_Student, ED_SoftLabelOnly) to RGB-domain benchmarks to probe their inductive biases. Our experiments show three robustness properties that distinguish these models from conventional CNNs trained only on RGB images. We summarize these main results in Figure 2 (overview in the right panel of Figure 1; comprehensive numbers in Appendices B–D).

4.1 Color and Hue Invariance

Standard CNNs are known to rely heavily on object color information for inference [1, 70, 71]. However, when we evaluate robustness using a dataset whose hue is rotated across the entire image,

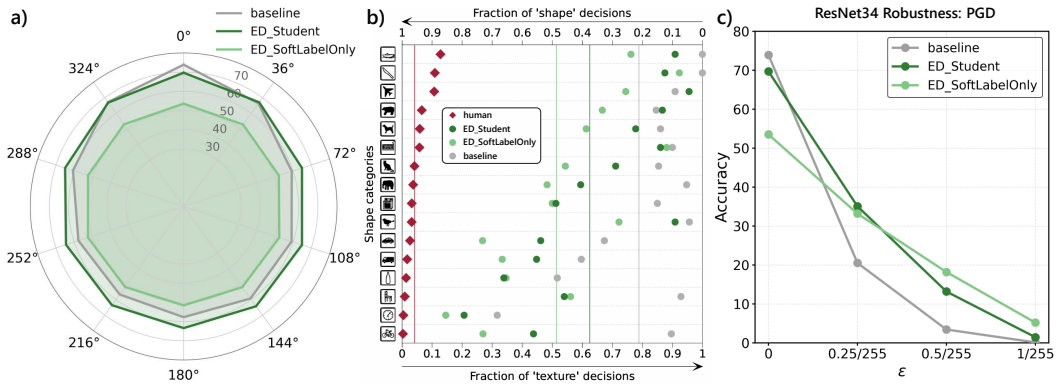


Figure 2: Robustness properties emerging from event-driven knowledge distillation. a) Top-1 accuracy across hue rotation angles. b) Cue-conflict shape bias; Event-distilled students approach the human reference. c) Accuracy under PGD attacks with $0 < \epsilon \leq 1/255$.

as shown in Figure 2 (a), the baseline model's accuracy drops sharply at certain rotation angles, whereas the Event-distilled model maintains color invariance across all angles. This color invariance is also confirmed when the ImageNet validation set is converted to grayscale (baseline: 73.9→62.4, ED_Student: 69.7→64.6; see Appendix B.2 Table 14), suggesting persistence against the loss of color information. Per-model comparisons are reported in Appendix B.

4.2 High Shape Bias

Next, to clarify what the Event-distilled model relies on for inference if not color, we performed the cue-conflict experiment proposed by Geirhos et al. (which classifies images that pit texture and shape information against each other) [1].

The scatter plot in Figure 2 (b) shows, for each category, the proportion of model decisions based on texture versus shape. While the baseline model, which performs texture-based inference, lies on the right side of the plot, the Event-distilled models (ED_Student and ED_SoftLabelOnly) shift substantially toward the left side and approach the region populated by humans, who make shape-based decisions. In other words, not only do the inference cues switch to shape, but the model also acquires shape-centric decision characteristics similar to those of humans.

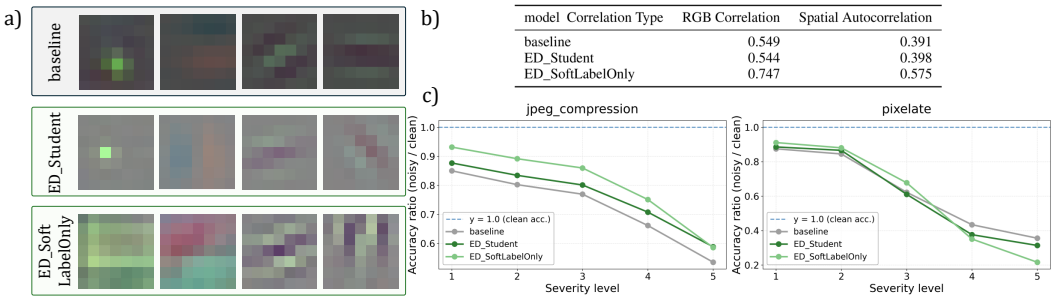
This shape bias is also corroborated by patch-shuffle experiments, in which the image is divided into a grid and shuffled [52, 72], and by evaluations under conditions in which color and texture information are stripped away to leave only shape (see Appendix C.3 Figure 12 and Appendix C.2 Table 14). As the mosaic grid becomes finer and the shape in the image is increasingly destroyed, the Event-distilled model's accuracy is observed to drop more sharply than the baseline's.

Two caveats apply to how we read these results. First, Burgert et al. [73] show that the preference measured by the cue-conflict experiment does not by itself imply strict reliance, so we describe our result as a shift of preference toward shape rather than a strict reduction of texture reliance. Second, their study systematizes controlled-suppression probes, of which our patch-shuffle and patch-rotation protocol is an instance, so we treat the two probes as complementary rather than interchangeable.

4.3 Robustness to High-Frequency Perturbations

The Event-distilled model also acquires resistance to deliberate noise and to adversarial perturbations on the input image.

In general, CNNs use high-frequency texture information as a key cue for classification, which makes them vulnerable to adversarial attacks such as FGSM [74] and PGD [75] that target small high-frequency components. As shown by the PGD evaluation in Figure 2 (c), the Event-distilled model attains accuracy above the baseline under perturbations with $0 < \epsilon \leq 1/255$, and the same trend holds for FGSM. Throughout this paper our adversarial claims are restricted to this perturbation range; we do not evaluate larger budgets or stronger attack ensembles such as AutoAttack [76]. Furthermore, in experiments that add band-limited random noise per frequency band, the



b)

model Correlation Type	RGB Correlation	Spatial Autocorrelation
baseline	0.549	0.391
ED_Student	0.544	0.398
ED_SoftLabelOnly	0.747	0.575

Figure 3: Internal-representation analysis and the spectral trade-off. a) First-layer filter excerpts, each min-max normalized to $[0, 1]$: baseline shows noisy high-frequency patterns; the Event-distilled model (especially ED_SoftLabelOnly) acquires smooth monochromatic, color-opponent, and Gabor-like structures. b) RGB and spatial autocorrelations; ED_SoftLabelOnly exceeds the baseline on both, while ED_Student stays close to it. c) Two ImageNet-C corruptions, as accuracy retention (noisy / clean). *Left*: under JPEG compression, the Event-distilled model retains more at high severity. *Right*: under pixelate, a reversal occurs at high severity.

Event-distilled model is confirmed to be more resistant than the baseline, especially against high-frequency-band noise (see Appendix D.2 Figure 15 for details of the frequency-band random-noise experiment). PGD evaluation hyperparameters are reported in Appendix D.1.

These three phenomena (color invariance, high shape bias, and resistance to high-frequency noise) suggest not a mere accuracy improvement but a shift in the representation space acquired by the model. In the next section, by analyzing the model's internal representation and its responses to real-world artifacts, we elucidate a unified mechanism that explains these seemingly disparate phenomena.

5 Mechanism: An Edge-Based Shape Feature Hypothesis

The three phenomena confirmed in the previous section (color invariance, high shape bias, and resistance to high-frequency noise) do not arise independently of one another. We hypothesize that knowledge distillation from event data causes the student model to suppress its dependence on unstructured high-frequency texture while strengthening its dependence on the macroscopic geometric structure that we call the edge-based shape feature. That is, whereas noise-augmentation methods broadly bolster post-hoc robustness against unstructured perturbations across all frequency bands, our distillation differs in that it selectively enhances dependence on the components that carry edge information. In fact, since event data record only brightness changes, the frequency components that correspond to edges in the original natural image are relatively emphasized, and as a result the radial PSD of N-ImageNet exhibits a smaller power-law exponent ($\alpha = 1.48$) than that of natural images (ImageNet, $\alpha = 2.46$); i.e., a flatter power spectral density (PSD) (see Appendix E for details). In this section, we examine this hypothesis through analysis of the model's internal representations and its responses to digital artifacts with different frequency characteristics.

5.1 Internal Representation: Structural Prior in First-Layer Filters

To elucidate the change in the internal representation acquired by the model, we analyzed the input-layer filters (Figure 3a). Whereas the baseline exhibits high-frequency noisy patterns that respond to local texture, the Event-distilled model shows high sensitivity to geometric structures. These filters can be broadly classified into (i) monochromatic filters that capture broad color distributions, (ii) color-opponent-like filters that capture macroscopic boundaries, and (iii) directional Gabor-filter-like edge-extraction structures (see Appendix C.4 for visualization of all 64 channels and detailed functional taxonomy). Importantly, these filters do not merely act as low-pass filters that smooth the input; rather, they acquire a structural prior that suppresses unstructured high-frequency texture while responding selectively to the structural gradients (edges) that constitute object contours. Appendix C.4.1 discusses why this is not an augmentation-pipeline artifact.

To quantitatively corroborate this change, we computed the channel-wise (RGB Correlation) and spatial (Spatial Autocorrelation) correlations of the filters (Figure 3b). ED_SoftLabelOnly shows substantially higher correlation values than baseline on both metrics; ED_Student remains close

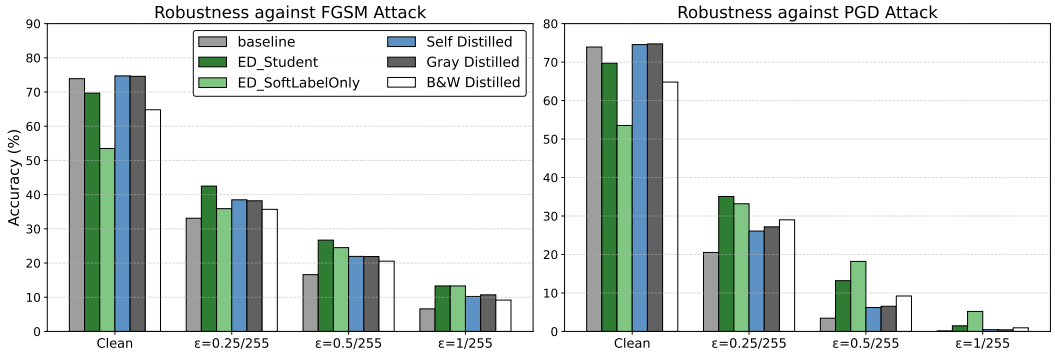


Figure 4: Ablation (Section 6.1). Adversarial robustness of color-stripped controls (Self / Gray / B&W Distilled) vs. the Event-distilled models. *Left*: FGSM. *Right*: PGD. x-axis: perturbation magnitude ϵ; y-axis: top-1 accuracy. The Event-distilled model surpasses all controls at $0 < \epsilon \leq 1/255$, indicating its inductive bias is not merely the result of color removal.

to baseline; since (b) aggregates over all filters while (a) shows individual ones, the two need not agree (Appendices C.4.1 and C.4.2). This indicates that, in ED_SoftLabelOnly, all channels synchronously extract geometric structures and the representation space has been transformed to emphasize continuity between adjacent pixels, which underlies the spectral trade-off discussed in the next subsection. How this property operates under actual color changes is corroborated by the early-layer feature-map visualization under hue rotation in Appendix B.3.

5.2 Validation via ImageNet-C: Spectral and Structural Trade-offs

To test the structural prior presented in the previous subsection, we evaluated all 19 corruption types of ImageNet-C. Robustness does not improve uniformly; it exhibits a spectral trade-off that depends on whether the corruption damages the frequency band or the geometric structure.

The Event-distilled model resists perturbations that selectively remove or occlude unstructured high-frequency texture. Under JPEG compression (Figure 3c, left) its retention exceeds the baseline's even at high severities, which suggests that stable inference is possible once shortcut texture is lost, as long as the macroscopic brightness gradients (edges) on which the model relies remain.

Conversely, it is vulnerable to perturbations that break the continuity of those geometric structures. Under pixelation (Figure 3c, right) it is resistant at low severities, where macroscopic contours are preserved, but at high severities, where edges become discontinuous, a reversal occurs and its retention falls below the baseline. Similar sensitivity appears under contrast and elastic_transform (all 19 corruptions are analyzed in Appendix D.3). This trade-off suggests that the decision boundary has shifted from superficial texture toward geometric structure and edge continuity.

5.3 Validation via Information Removal: Dominance of Geometric Features

To test the edge-based shape feature hypothesis more directly, we evaluate inference under shape-only conditions in which the ImageNet validation set is converted to B&W (Otsu thresholding [77]), Canny edges [78], or silhouette. Under these conditions the baseline degrades sharply; ED_Student also degrades but stays above the baseline on all three (see Appendix C.2 for the per-model table).

Together with the smooth edge-extracting filters of Section 5.1, these results indicate that the inference basis of the Event-distilled model has shifted from texture to geometry (shape).

6 Validation and Practical Implications

The previous sections established that the Event-distilled model acquires an edge-based shape feature. Here we validate it with ablations, compare it against existing robustification techniques, and assess its practical value through Linear-probe transfer on downstream tasks.

6.1 Ablation Study: Is it merely the absence of color?

To examine whether the high shape bias and robustness acquired by the Event-distilled model are due to the mere absence of color information, we prepared distillation models from teacher models

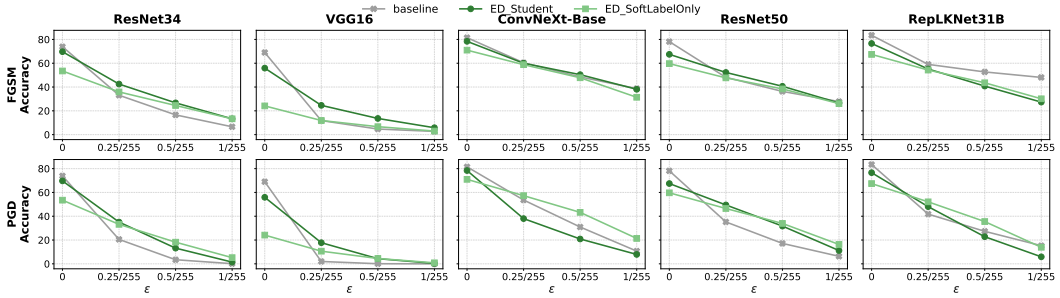


Figure 5: Architectural generality (Section 6.2). Comparison of baseline / ED_Student / ED_SoftLabelOnly across ResNet-34, VGG-16, ConvNeXt-Base, ResNet-50, and RepLKNet31B [79]. Top: FGSM. Bottom: PGD. x-axis: perturbation magnitude ϵ. ED_Student improves over the baseline in 19 of the 30 tested (architecture, attack, ϵ) settings with $0 < \epsilon \leq 1/255$. The 11 exceptions, reported as the ED_Student − baseline difference in percentage points, are (RepLKNet31B, FGSM) at $\epsilon = 0.25, 0.5, 1/255$: -3.9, -11.9, -20.7; (RepLKNet31B, PGD) at $\epsilon = 0.5, 1/255$: -4.5, -9.4; (ConvNeXt-Base, PGD) at $\epsilon = 0.25, 0.5, 1/255$: -15.6, -10.1, -2.8; (ConvNeXt-Base, FGSM) at $\epsilon = 0.25, 1/255$: -0.4, -0.6; and (ResNet-50, FGSM) at $\epsilon = 1/255$: -1.2.

trained on ImageNet RGB images, grayscale images, and fully color-stripped binary (B&W) images, and conducted comparative validation against the corresponding distilled students (Self Distilled, Gray Distilled, B&W Distilled).

Under adversarial attack (Figure 4), the Event-distilled model resists better than all controls. Its superiority even over B&W Distilled suggests that the transferred bias is not what one obtains by forcing the student to learn contours, but a representation rooted in sensitivity to the brightness gradients characteristic of the event sensor.

6.2 Architectural Generality

Next, we verified whether the phenomena reported in the main paper depend on a particular network structure. In addition to the ResNet-34-based validation, we performed the same Event distillation and evaluation using VGG-16 (a different structure), the deeper ResNet-50, the more recent ConvNeXt-Base, and RepLKNet31B [79], a CNN with 31×31 large kernels, and examined two axes (high-frequency noise robustness, Section 4.3; and shape bias, Section 4.2) across these architectures.

High-frequency noise robustness. As shown in Figure 5, ED_Student improves over the baseline in 19 of the 30 tested (architecture, attack, ϵ) settings with $0 < \epsilon \leq 1/255$ under FGSM and PGD (5 architectures $\times$ 2 attacks $\times$ 3 values of $\epsilon \in \{0.25, 0.5, 1\}/255$). The 11 exceptions are enumerated in the caption of Figure 5; they are concentrated in the two largest backbones, ConvNeXt-Base and RepLKNet31B.

Shape bias. Both the cue-conflict experiment (Appendix C.1) and the patch-shuffle / patch-rotation experiment that destroys macroscopic shape continuity (Appendix C.3) consistently reproduce, across all architectures, the behavior observed for ResNet-34: a shift toward the shape side and a steeper accuracy drop as the grid becomes finer.

These cross-model evaluations suggest that the cross-domain distillation we propose injects its inductive bias without depending on a particular network structure, and that the main benefits reported in the paper, shape bias and high-frequency noise robustness, are largely insensitive to architectural design choices, with the exceptions noted above.

6.3 Transferability: Event Distillation as a Structural Prior

An additional finding regarding application is that the inductive bias acquired by the Event-distilled model functions as a useful prior in transfer learning to downstream tasks (Linear probe [80]). We perform Linear-probe evaluations on 19 image-classification datasets to compare the transfer performance of ED_Student against baseline. As points of comparison, we prepare representative robustification techniques (Adversarial Training, PGD-based [75], and Noise Training, Gaussian noise injection [81]) as well as the SIN+IN co-training model (the SIN model) [1] that directly induces

	C100	C10	ChXR	MNIST	Cars	DTD	STL10	INet-sk	Cal101	INet-v2	Scene	SUN397	Flwr	KITTI	ESAT	CUB	PCam	Pet	Food	Mean	INet
base. (%)	55.8	78.0	87.3	89.9	46.9	60.1	94.8	62.2	85.4	49.6	90.6	53.5	85.3	73.4	90.0	64.6	78.0	90.5	**61.4**	73.5	73.9
ED (Δ)	5.2	4.8	**4.3**	3.4	**2.3**	**1.9**	**1.8**	1.4	**1.2**	0.9	**0.8**	**0.8**	**0.7**	0.7	**0.4**	**0.3**	0.2	−0.1	−0.4	**1.6**	−4.2
Adv. (Δ)	**14.6**	**10.8**	2.9	**5.6**	−8.7	−9.1	−0.7	−5.2	0.4	−12.3	−0.5	−7.8	−9.4	2.1	−1.2	−3.4	0.0	−3.2	−9.5	−1.8	−7.9
Noise (Δ)	3.2	2.6	0.6	−0.7	−5.6	−2.5	0.6	−4.5	0.7	1.7	−0.8	−2.3	−3.9	0.4	−1.7	0.2	**0.3**	−0.1	−3.8	−0.8	**0.5**
SIN (Δ)	6.2	5.6	0.0	1.8	−8.3	−4.2	1.4	−4.5	0.7	−4.1	−0.3	−1.9	−5.5	**3.2**	−2.2	−3.0	−1.2	−0.6	−1.1	−0.9	−1.7
Canny (Δ)	−1.8	1.0	0.2	0.3	1.1	1.4	0.9	**1.5**	−0.1	**2.9**	0.2	0.2	0.1	−0.1	−1.7	−2.1	0.3	**0.1**	−0.7	0.2	−2.8
Synth (Δ)	−9.4	−6.0	−2.1	−5.0	−4.2	−6.7	−0.1	−4.9	−3.8	−2.4	−2.3	−11.7	−0.9	−0.9	−7.2	−21.8	−0.6	−1.0	−12.9	−5.4	−2.1

Table 1: Linear probe transfer (Sections 6.3, 6.4). base. gives baseline top-1 (%); the Δ rows (%, + omitted) are ED_Student, Adversarial Training, Noise Training, SIN+IN co-training [1], Canny_Distilled, and Synth_Distilled. Mean averages the 19 transfer tasks; INet is the pre-training reference, excluded from Mean. Bold: column-wise best; datasets sorted by ED_Student's Δ; abbreviations expanded in Appendix A.5.

	Clean	Δ Color (hue Avg.)	Δ Shape (sketch)	Δ High-freq (bandpass)	Δ Transfer (LP Avg.)	Axes improved
baseline (abs.)	73.9	61.8	23.7	29.6	73.5	—
ED_Student	69.7	+3.6	+4.4	+17.4	+1.6	**4 / 4**
Canny_Distilled	71.1	+4.3	+4.2	−5.9	+0.2	3 / 4
Synth_Distilled	71.8	−3.4	−3.2	+18.9	−5.4	1 / 4
Adv. Training	66.0	−11.2	−2.4	+28.1	−1.8	1 / 4
Noise Training	74.4	+0.8	+1.7	+13.4	−0.8	3 / 4
SIN	72.2	+0.5	+7.1	+19.3	−0.9	3 / 4

Table 2: Event supervision vs. five non-event supervisions, all with the same ResNet-34 student. Clean is absolute for every row; on the four axes the baseline row is absolute and the others are differences from it, signed so that positive means better. Color: mean top-1 over ten hue-rotation angles; Shape: ImageNet-Sketch top-1 [85]; High-freq: reduction in the accuracy drop under band-limited noise at 0.95π, where the baseline drops 29.6 pp; Transfer: mean linear-probe accuracy over the 19 tasks of Table 1. Only ED_Student improves on all four axes.

shape bias, and report the performance gain over the baseline (Δ Score) in Table 1. Adversarial Training and Noise Training are reported to enhance robustness against adversarial noise and blur while degrading markedly under perturbations such as fog [81], in contrast to the Event-distilled model.

Notably, although the Event-distilled model is 4.2 percentage points below the baseline on pre-training ImageNet accuracy, it outperforms the baseline on many downstream tasks under Linear-probe transfer. The largest gains appear on CIFAR-100 / CIFAR-10 [82], Chest X-ray [83] and MNIST [84], and gains are also observed where texture is largely absent or fine-grained geometry carries the signal, such as imagenet-sketch [85] and cars196 [86]; per-dataset values are in Table 1. We attribute this to the generality of the representation space acquired during pretraining: by suppressing dataset-specific high-frequency texture and selectively extracting geometric structure, the model acts as a structural prior for unseen tasks under the frozen-weights Linear-probe setting. Hyperparameters and protocol details are in Appendix A.5.

6.4 Necessity: What Non-Event Supervisions Fail to Reproduce

Is event data required at all for the properties reported above? We compare ED_Student against five non-event supervisions, all with the same ResNet-34 student, on four axes: the three phenomena of Section 4 together with the downstream transfer of Section 6.3. Two of the five are new here: Canny_Distilled, whose ResNet-34 teacher is trained from scratch on Canny edge maps of ImageNet (teacher 55.82%, student 71.15%), representing RGB-derived edge and gradient operators; and Synth_Distilled, whose teacher is trained on a two-channel ON/OFF proxy formed from the grayscale difference between two augmented views of the same image (teacher 64.52%, student 71.78%), representing proxies that imitate event data without an event sensor. The distillation controls follow the ED_Student recipe; the robustification baselines keep their own (Table 6). The color-stripped controls of Section 6.1 are further non-event supervisions, not repeated here; generation and training details are in Appendix A.8. Table 2 summarizes the resulting four-axis profiles.

Canny_Distilled reproduces the color and shape aspects: hue rotation (66.1) and ImageNet-Sketch (27.9) are at ED_Student's level. It nevertheless fails on high-frequency-band robustness, dropping 35.5 pp at 0.95π against the baseline's 29.6 pp, and its downstream gain ($+0.2$ pp) does not reach ED_Student's ($+1.6$ pp). Its cue-conflict shape bias (0.270) also stays closer to the baseline (0.212) than to ED_Student (0.375): the shape transfer it achieves shows up in sketch-style inputs but not in the shape-versus-texture preference.

Synth_Distilled shows the complementary pattern: it matches or exceeds ED_Student on the frequency and adversarial axes (bandpass drop 10.7 pp; PGD accuracy-curve area 10.2 against 8.9, baseline 6.2), yet falls below the baseline on hue rotation (58.4), ImageNet-Sketch (20.5), and downstream transfer (-5.4 pp). Adversarial Training points the same way: it attains the largest PGD accuracy-curve area (24.1) and the smallest bandpass drop (1.5 pp) while losing 11 pp on hue rotation. The adversarial and high-frequency gains are therefore not uniquely attributable to event supervision.

Existing robustification methods are partial in the same way. Adversarial and Noise Training confer targeted resistance to the perturbations they are trained against, but robustness and clean accuracy are known to be at odds [87], and both fall below the baseline in linear-probe transfer. SIN attains the highest accuracy of all models once color and texture are removed (Table 14), yet falls below the baseline on 12 of 19 downstream tasks; since the cue-conflict benchmark is produced by the same style-transfer procedure as Stylized-ImageNet, SIN is evaluated close to its own training distribution and its advantage does not carry over to ordinary RGB images.

Across the six models, ED_Student is the only one that improves over the baseline on all four axes at once. We therefore do not claim that any single property requires event data; rather, it is the joint profile that the non-event supervisions examined here fail to reproduce.

Do the gains survive a stronger baseline? Retraining both the baseline and ED_Student under a timm-style recipe (300 epochs with stochastic depth, label smoothing, MixUp, and RandAugment) reaches 75.9% and 70.8% clean accuracy. ED's advantage persists on shape ($+2.0$ pp), color ($+1.8$ pp), high-frequency-band robustness (-4.2 pp drop), and downstream transfer ($+1.3$ pp), but not on the adversarial axis, where the stronger augmentation already makes the SOTA baseline robust at these perturbation sizes. Full numbers are in Appendix A.9.

7 Limitations and Future Work

N-ImageNet is produced by recapturing RGB images on a monitor, so the sensor's high dynamic range is not exercised and the planar source has no motion parallax; our claims are restricted to the brightness-gradient selectivity of event data. The effect also appears to diminish with model capacity: ConvNeXt-Base and RepLKNet31B gain less over their baselines than ResNet-34 and account for most of the 11 exceptions in Figure 5. Our evaluation is confined to CNNs and image classification. Three directions follow: evaluation on a broader range of vision models beyond CNNs, including Vision Transformers [88] and dense-prediction tasks; comparison against feature-level distillation such as FitNets [58], since our loss operates on logits alone; and using Eq. 1 as a fine-tuning objective.

8 Conclusion

To elucidate the inductive bias that event data brings to vision models, we proposed an analysis framework based on cross-domain knowledge distillation. The Event-distilled model suppresses the unstructured high-frequency texture dependence of conventional CNNs and instead depends strongly on the edge-based shape feature. We characterize this as the first systematic empirical study separating what event-domain supervision transfers into RGB representations from what RGB-based alternatives also achieve under a common student architecture. The mechanism manifests as smoother early-layer filters and a spectral trade-off: robustness to texture loss coexists with vulnerability to frequency-band contamination and geometric disruption. Beyond serving as an alternative sensor output, event data can be a source of knowledge for more general-purpose visual representations.

Acknowledgments and Disclosure of Funding

MT was partially supported by JSPS KAKENHI (22H05116), JST CREST (JPMJCR22N4), and AMED under Grant Number JP25wm0625422.

References

[1] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1811.12231.

[2] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (1):154–180, 2022. URL https://arxiv.org/abs/1904.08405.

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.12261.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

[5] Junho Kim, Jaehyeok Bae, Gangin Park, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. *CoRR*, abs/2112.01041, 2021. URL https://arxiv.org/abs/2112.01041.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. URL https://arxiv.org/abs/1409.0575.

[7] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[8] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. Poker-dvs and mnist-dvs. their history, how they were made, and other details. *Frontiers in neuroscience*, 9:481, 2015.

[9] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405, 2016.

[10] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:244131, 2017.

[11] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.

[12] Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017.

[13] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018.

[14] Diederik Paul Moeys, Daniel Neil, Federico Corradi, Emmett Kerr, Philip Vance, Gautham Das, Sonya A Coleman, Thomas M McGinnity, Dermot Kerr, and Tobi Delbruck. Pred18: Dataset and further experiments with davis event camera in predator-prey robot chasing. *arXiv preprint arXiv:1807.03128*, 2018.

[15] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 491–501, 2019.

[16] Wensheng Cheng, Hao Luo, Wen Yang, Lei Yu, Shoushun Chen, and Wei Li. Det: A high-resolution dvs dataset for lane extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[17] Ajay Vasudevan, Pablo Negri, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, pages 675–682. IEEE, 2020.

[18] Bharath Ramesh, Andrés Ussa, Luca Della Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in neuroscience*, 14:135, 2020.

[19] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022.

[20] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19866–19877, 2023.

[21] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023.

[22] Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2378–2388, 2024.

[23] Jiahang Cao, Mingyuan Sun, Ziqing Wang, Hao Cheng, Qiang Zhang, Shibo Zhou, and Renjing Xu. Spiking neural network as adaptive event stream slicer. *Advances in Neural Information Processing Systems*, 37:75064–75094, 2024.

[24] Yan Yang, Liyuan Pan, Dongxu Li, and Liu Liu. Ezsr: Event-based zero-shot recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4628–4638, 2025.

[25] Marcin Kowalczyk, Kamil Jeziorek, and Tomasz Kryjak. Learning from noise: Enhancing dnns for event-based vision through controlled noise injection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5092–5102, 2025.

[26] Rui Fan, Weidong Hao, Juntao Guan, Lai Rui, Lin Gu, Tong Wu, Fanhong Zeng, and Zhangming Zhu. Eventpillars: Pillar-based efficient representations for event data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2861–2869, 2025.

[27] Quanmin Liang, Qiang Li, Shuai Liu, Xinzi Cao, Jinyi Lu, Feidiao Yang, Wei Zhang, Kai Huang, and Yonghong Tian. Efficient event camera data pretraining with adaptive prompt fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8656–8667, 2025.

[28] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *European Conference on Computer Vision*, pages 477–494. Springer, 2024.

[29] Wenhao Xu, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Ceia: Clip-based event-image alignment for open-world event-based understanding. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[30] Sungheon Jeong, Hanning Chen, Sanggeon Yun, Suhyeon Cho, Wenjun Huang, Xiangjian Liu, and Mohsen Imani. Cross-modal event encoder: Bridging image-text knowledge to event streams. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3213–3222, 2026.

[31] Weiqi Yan, Chenlu Lin, Youbiao Wang, Zhipeng Cai, Xiuhong Lin, Yangyang Shi, Weiquan Liu, and Yu Zang. Omnievent: Unified event representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11568–11576, 2026.

[32] Ruichen Ma, Liwei Meng, Guanchao Qiao, Ning Ning, Yang Liu, and Shaogang Hu. I2e: Real-time image-to-event conversion for high-performance spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1982–1990, 2026.

[33] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

[34] Anindya Mondal, Jhony H Giraldo, Thierry Bouwmans, Ananda S Chowdhury, et al. Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 876–884, 2021.

[35] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023.

[36] Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12846–12856, 2023.

[37] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629 (8014):1034–1040, 2024.

[38] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5819–5828, 2024.

[39] Nan Yang, Yang Wang, Zhanwen Liu, Meng Li, Yisheng An, and Xiangmo Zhao. Smamba: Sparse mamba for event-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9229–9237, 2025.

[40] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[41] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022.

[42] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023.

[43] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15686–15698, 2024.

[44] Hebei Li, Yansong Peng, Jiahui Yuan, Peixi Wu, Jin Wang, Yueyi Zhang, and Xiaoyan Sun. Efficient event-based semantic segmentation via exploiting frame-event fusion: A hybrid neural network approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18296–18304, 2025.

[45] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 608–619, 2021.

[46] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10699–10709, 2023.

[47] Jinchang Zhang, Zijun Li, Jiakai Lin, and Guoyu Lu. Adaptive event stream slicing for open-vocabulary event-based object detection via vision-language knowledge distillation. *arXiv preprint arXiv:2510.00681*, 2025.

[48] Peng Wu, Yuting Yan, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. Towards video anomaly detection from event streams: A baseline and benchmark datasets. *arXiv preprint arXiv:2603.24991*, 2026.

[49] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49 (10):2333–2341, 2014.

[50] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12):e1006613, 2018.

[51] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[52] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations (ICLR)*, 2019.

[53] Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[54] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in cnns. In *International Conference on Learning Representations (ICLR)*, 2021.

[55] Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *International Conference on Learning Representations (ICLR)*, 2021.

[56] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations (ICLR)*, 2021.

[57] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34*, 2021.

[58] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

[59] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021.

[62] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[63] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[64] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[65] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[66] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision (ECCV)*, 2018.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1512.03385.

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.1556.

[69] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2201.03545.

[70] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[71] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019.

[72] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[73] Tom Burgert, Oliver Stoll, Paolo Rota, and Begüm Demir. Imagenet-trained cnns are not biased towards texture: Revisiting feature reliance through controlled suppression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[74] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

[75] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[76] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 2206–2216, 2020.

[77] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[78] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, PAMI-8(6):679–698, 1986.

[79] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2203.06717.

[80] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017. URL https://arxiv.org/abs/1610.01644.

[81] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[82] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[83] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[84] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[85] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1905.13549.

[86] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013. URL https://ai.stanford.edu/~jkrause/papers/3drr13.pdf.

[87] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[88] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

[89] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

[90] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proc. International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1608.03983.

[91] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://arxiv.org/abs/1905.04899.

[92] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.09412.

[93] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. URL https://arxiv.org/abs/1909.13719.

[94] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proc. AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/abs/1708.04896.

[95] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[97] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

[98] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2202.06709.

[99] A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996.

[100] Michael S. Langer. Large-scale failures of $f^{-\alpha}$ scaling in natural image spectra. *Journal of the Optical Society of America A*, 17(1):28–33, 2000.

A Experimental Setup

A.1 Datasets

In this work, we adopt Neuromorphic-ImageNet (N-ImageNet) [5] as the dataset for the event domain, and ILSVRC2012 (ImageNet) as the dataset for the RGB domain. N-ImageNet has the same 1000-class structure as ImageNet and is generated by displaying each image of ImageNet on a monitor and recapturing it with a DVS-type event camera. As such, every event sample in N-ImageNet is paired at the instance level with the corresponding RGB image in ImageNet. The Stylized-ImageNet dataset [1], used in our SIN comparison (Section 6.4), is likewise a 1000-class style-transferred variant of ImageNet with the same 1,281,167 training images.

The splits used for training and evaluation are as follows.

- Teacher training (Section A.2): Only the training split of N-ImageNet is used for training, and the validation split is used for verification.
- Student training (Section A.3): Strongly augmented RGB inputs from the ImageNet training split are used to compute the CE loss, while the events from the N-ImageNet training split corresponding to the same instance are passed through the teacher model to compute the KL loss. Validation is performed on the ImageNet validation split with top-1/top-5 accuracy.

A.2 Teacher Training

A.2.1 Event Teacher (DiST ResNet-34)

As the Event teacher in this work, we directly use the DiST ResNet-34 checkpoint released by the authors of the prior work [5], in a frozen state. This model is a ResNet-34 that takes the DiST representation as input, and it achieves 48.43% top-1 accuracy on the N-ImageNet validation split.

In all Event-distillation experiments with ResNet-34 (ED_Student, ED_SoftLabelOnly, and the transfer experiments from Section 6.3 onward), we use these frozen weights as a common teacher.

A.2.2 Teacher Models for Comparison

The control teachers used in the ablation of Section 6.1 are all ResNet-34 variants trained on ImageNet with the CE loss.

- Teacher for Self Distilled: A ResNet-34 baseline trained normally on RGB ImageNet is used as the teacher.
- Teacher for Gray Distilled: A ResNet-34 trained on the ImageNet RGB images converted to 3-channel grayscale.
- Teacher for B&W Distilled: A ResNet-34 trained on ImageNet preprocessed by OpenCV's Otsu thresholding [77] (cv2.THRESH_BINARY | cv2.THRESH_OTSU). To preserve the binary property, all spatial transforms use NEAREST interpolation (see Section A.4.3).

These three teachers are all trained for 120 epochs using AdamW [89] (learning rate $= 2 \times 10^{-3}$, weight decay $= 5 \times 10^{-2}$), with 5 epochs of linear warmup followed by cosine scheduling [90] that decays to 0 over the remaining epochs. CutMix and MixUp are not applied.

A.2.3 Additional Teacher Models for Architectural Generality

In the architectural-generality validation of Section 6.2, we additionally trained DiST teachers in-house for 4 architectures other than ResNet-34 (ResNet-50, VGG16-BN, ConvNeXt-Base, RepLKNet31B). Note that the DiST ResNet-34 teacher [5] of Section A.2.1 uses the publicly released weights as-is, and is therefore not part of the additional training in Section 6.2.

The training recipe shared across the four architectures (input representation, event slicing, batch size, weight decay, momentum, number of epochs, parallelization scheme) is shown in Table 3, while the architecture-specific settings (optimizer, learning rate, first-conv kernel size) are shown in Table 4. CutMix and MixUp are enabled only when training the RepLKNet31B teacher (see Section A.4.6).

Item	Value
Input representation	DiST (`reshape_then_acc_adj_sort`), 2 ch, 224×224
Event slicing	Random, `slice_length` $= 30{,}000$
Batch size	512
Weight decay	1×10^{-4}
Momentum	0.9
Epochs	100
Parallelism	DataParallel

Table 3: Shared training recipe for the in-house DiST teachers of Section 6.2. Architecture-specific settings are in Table 4. The DiST ResNet-34 teacher used elsewhere is the publicly released checkpoint of [5] and is not retrained by us.

A.3 Student Training

This subsection describes in detail the training settings for the student models in the main paper. All settings below pertain to the ResNet-34 student.

Architecture	Optimizer	Learning rate	First-conv kernel size
ResNet-50	Adam	2×10^{-4}	14
VGG16-BN	Adam	2×10^{-4}	14
ConvNeXt-Base	AdamW	2×10^{-4}	4
RepLKNet31B	AdamW	2×10^{-4}	4

Table 4: Architecture-specific settings for the DiST teachers in Section 6.2. Common settings are in Table 3. The DiST ResNet-34 teacher uses the public checkpoint of [5] and is omitted.

A.3.1 Loss Function

All KD students minimize the composite loss given in Eq. 1 of the main paper. The mixing weight α corresponding to each model and the teacher used are shown in Table 5. The temperature in $\mathcal{L}_{\mathrm{KL}}$ is fixed at $T = 3.0$ across all KD students.

For the Adversarial Training and Noise Training models, the KL term is not used (corresponding to $\alpha = 0$); only a cross-entropy loss with label smoothing $\varepsilon = 0.1$ is minimized. The SIN model is trained on the union of ImageNet and Stylized-ImageNet with a plain (non-smoothed) cross-entropy loss only (see Section A.3.4).

Model	α	Teacher
baseline	0.0	–
ED_Student (main)	0.2	DiST ResNet-34
ED_SoftLabelOnly	1.0	DiST ResNet-34
Self Distilled	0.2	RGB ResNet-34 baseline
Gray Distilled	0.2	Grayscale ResNet-34
B&W Distilled	0.2	B&W ResNet-34

Table 5: Distillation weight α (Eq. 1) and the teacher used for each ResNet-34 student. The KL-term temperature is fixed at $T = 3.0$. Adversarial / Noise Training and the SIN model use only a CE loss and are omitted from this table (see Section A.3.3, A.3.4, respectively).

A.3.2 Optimization and Learning Rate Scheduling

For all ResNet-34 student models, we adopt a batch size of 512, AdamW as the optimizer, a common random seed (seed=1), and DDP parallelization. Per-model learning rates, numbers of epochs, and learning-rate schedules are shown in Table 6.

Model	LR	Weight decay	Epochs	LR schedule
baseline / ED_Student / ED_SoftLabelOnly	2×10^{-3}	5×10^{-2}	120	5-epoch linear warmup $\rightarrow$ cosine
Self / Gray / B&W Distilled	7.5×10^{-4}	5×10^{-2}	120	5-epoch linear warmup $\rightarrow$ cosine
SIN	2×10^{-3}	5×10^{-2}	120	5-epoch linear warmup $\rightarrow$ cosine
Adversarial Training	1×10^{-3}	2×10^{-4}	50	Cosine annealing (no warmup)
Noise Training	1×10^{-3}	2×10^{-4}	120	Cosine annealing (no warmup)

Table 6: Per-student optimization configuration for the ResNet-34 student models.

A.3.3 Additional Settings for Adversarial Learning/Noise Learning

- PGD attack (Adversarial Training): perturbation bound $\epsilon = 8/255$, step size $\alpha = 2/255$, number of iterations $K = 10$. For each batch, after initializing with uniform random noise of magnitude ϵ, we perform K signed-gradient updates and clip back to within ϵ in the L_∞ norm at the end of each step. The student is trained only on the generated adversarial inputs, and validation is performed on clean images.
- Gaussian Noise Injection (Noise Training): Gaussian noise with mean 0 and standard deviation $\sigma = 0.1$ is added to the tensor image with probability $p = 0.5$, and the result is

clipped to $[0, 1]$. This layer is inserted immediately before Normalize. No noise is added at validation.

A.3.4 Additional Settings for SIN+IN Co-training

Following the SIN+IN variant of Geirhos et al. [1], the SIN model is trained as a single ResNet-34 on a ConcatDataset that joins the training splits of ImageNet and Stylized-ImageNet (2,562,334 images in total). Each epoch draws samples from the two datasets in a 1:1 ratio. Both datasets share the same 1000-class WNID set as ImageNet, and we assert in the loader that the two share the same $class_to_idx$. Validation is performed on the ImageNet validation split.

The KD term is not used and label smoothing is not applied; only a plain cross-entropy loss is minimized. The optimizer, schedule, seed, and parallelization are kept identical to baseline / ED_Student (Table 6), so that no confound other than the dataset is introduced. The augmentation pipeline is also the same RGB strong pipeline (Table 7) applied identically to both datasets. Trained for 120 epochs under this setup, the model achieves a top-1 accuracy of 72.18% on the ImageNet validation split.

A.4 Data Augmentation

A.4.1 RGB pipelines (students and RGB-based teachers)

We use two RGB augmentation pipelines, chosen by purpose:

- a **strong** pipeline on the CE branch of every student (Table 7), and
- a **weak** pipeline on the KD branch and on every validation loader (Table 8).

RandAugment is applied to every batch, i.e. with probability $p = 1.0$. Adversarial Training uses a simplified strong pipeline that omits RandAugment and RandomErasing, so that the augmentations do not interfere with the PGD perturbations. Noise Training keeps the full strong pipeline and additionally inserts the Gaussian-noise layer described in Section A.3.3 before Normalize.

A.4.2 Grayscale pipelines

For the Grayscale teacher and for the KD branch of Gray Distilled, we extend the RGB pipelines of Section A.4.1 by inserting `torchvision.transforms.v2.Grayscale(num_output_channels=3)` (i.e. after all spatial and RandAugment operations). The ImageNet normalization statistics (mean and standard deviation) are reused unchanged, applied to the three-times replicated luminance channel.

A.4.3 B&W pipelines

Binarized inputs are used by the B&W teacher and by the KD branch of B&W Distilled. Preserving the two-valued pixel distribution is critical here, so we (i) use NEAREST interpolation for every spatial operation, and (ii) drop the pixel-smoothing augmentations RandAugment and RandomErasing. The resulting pipeline is listed in Table 9.

The B&W images themselves are produced offline by OpenCV's Otsu thresholding (`cv2.THRESH_BINARY | cv2.THRESH_OTSU`). A fallback path applies the same thresholding on the fly whenever the precomputed image is missing from disk.

A.4.4 Event-side augmentation

Event inputs use only three operations: random time flip, random horizontal flip, and random spatial shift (up to ± 20 pixels). No counterpart to the strong RGB pipeline, i.e. color- or texture-based augmentations such as RandAugment or RandomErasing, is applied on the event side. These three event-side operations are sampled independently per iteration during distillation, matching the recipe used to train the original DiST teacher [5].

A.4.5 Cross-Modal Augmentation Alignment in the KD Branch

A consequence of the pipelines described above is that, at distillation time, the teacher's event input x_{event} is passed through the stochastic event-side pipeline of Section A.4.4 (random time flip, random

horizontal flip, ± 20 px spatial shift), whereas the student's KD branch consumes a deterministic Resize/CenterCrop view x_{weak} of the paired RGB image (Table 8). The two views of a given paired instance are therefore not synchronized on a per-iteration basis. We argue below that this asymmetry does not introduce systematic label noise into the distillation target.

(A) The teacher–student view asymmetry follows established KD practice. Asymmetric augmentation between the teacher and student branches is standard in modern distillation. DeiT [61] feeds the teacher with weakly augmented inputs while applying strong augmentations (RandAugment, MixUp, CutMix) to the student; Patient Teacher [62] studies teacher–student view consistency in distillation and reports that feeding the teacher and the student the same view is beneficial. Our cross-modal setting cannot share a view, since the two branches consume different modalities; the resulting view mismatch is a confound that we did not isolate experimentally.

(B) The KL is computed in class-probability space; spatial misalignment is not pixel-level label noise. The teacher's event-side augmentations (random horizontal flip, ± 20 px spatial shift) preserve the class identity of the underlying instance: a flipped or slightly shifted view of a butterfly event is still a butterfly. The KL divergence in Eq. 1 is computed between class-probability distributions p_t and p_s [4], not between pixel-aligned predictions. Hence the teacher–student view mismatch translates into a stochastic regularization signal (the student is asked to match the teacher's class distribution across slightly perturbed views of the same instance) rather than into systematic label noise. This is analogous to consistency regularization in self-supervised learning, where two augmented views of the same image are pulled together in representation space.

(C) Empirical coherence of behavioral signatures is inconsistent with systematic noise. If the cross-modal view asymmetry produced systematic label noise that distorted the distillation target, we would expect ED_Student / ED_SoftLabelOnly to exhibit degraded or random behavior. Instead, both models display a coherent and trackable set of signatures aligned with the edge-based shape feature hypothesis: color invariance (Section 4.1), shape bias (Section 4.2), high-frequency-noise robustness (Section 4.3), the JPEG / pixelate trade-off (Section 5.2), shape-only inference (Section 5.3), and downstream transfer (Section 6.3). Stochastic label noise would not yield such a directional and reproducible representation shift; the empirical evidence is therefore consistent with the view-mismatch acting as a benign regularizer rather than as a source of systematic distortion.

A.4.6 Scope of CutMix and MixUp

CutMix [91] and MixUp [92] are used only when training the RepLKNet31B teacher and its corresponding student. When both are enabled, one of the two is selected uniformly at random ($p = 0.5$) per mini-batch. For every other model (the ResNet-34, ResNet-50, VGG16-BN, and ConvNeXt-Base teachers, as well as every ResNet-34 student studied in this paper), neither CutMix nor MixUp is applied, except for the stronger-recipe runs of Appendix A.9, which use MixUp.

Operation	Setting
RandomResizedCrop	output 224, scale $= (0.08, 1.0)$
RandomHorizontalFlip	$p = 0.5$
RandAugment	`num_ops` $= 2$, `magnitude` $= 9$
Normalize	ImageNet mean/std
RandomErasing	$p = 0.25$

Table 7: RGB strong augmentation pipeline used on the CE branch of every ResNet-34 student. Adversarial Training omits RandAugment [93] and RandomErasing [94] to avoid interference with PGD noise.

Operation	Setting
Resize	256
CenterCrop	224
Normalize	ImageNet mean/std

Table 8: RGB weak augmentation pipeline used on the KD branch (student input during KD loss computation) and on every validation loader.

Operation	Setting
RandomResizedCrop	output 224, scale $= (0.08, 1.0)$, **NEAREST** interpolation
RandomHorizontalFlip	$p = 0.5$
Normalize	ImageNet mean/std

Table 9: Augmentation pipeline for binarized (B&W) inputs. All spatial operations use NEAREST interpolation to preserve the two-valued pixel distribution; RandAugment and RandomErasing are omitted.

A.5 Downstream Linear-Probe Protocol and Hyperparameters

The downstream linear-probe evaluation reported in Section 6.3 of the main paper follows a strict linear-probe protocol: the entire ResNet-34 backbone is frozen, and only a single newly initialized linear layer (`nn.Linear(in_features, num_classes)`) is trained on each downstream dataset. This isolates the transferability of the pretrained features and prevents the backbone capacity from confounding the comparison between models.

Training hyperparameters are kept identical across all five compared models (baseline, ED_Student, Adversarial Training, Noise Training, SIN) and across all 19 downstream datasets:

- Optimizer: SGD with momentum 0.9, weight decay 1×10^{-4}.
- Learning rate: 0.1 with cosine annealing (`CosineAnnealingLR`, $T_{\max} =$ epochs).
- Epochs: 50, batch size: 1024, loss: cross-entropy without label smoothing.
- Train transform: `RandomResizedCrop`(224, scale 0.08–1.0) + `RandomHorizontalFlip` + `RandAugment`(`num_ops`$=2$, `magnitude`$=9$) + ImageNet `Normalize` + `RandomErasing`($p = 0.25$).
- Validation transform: `Resize`(256, antialias) + `CenterCrop`(224) + ImageNet `Normalize`.
- Two seeds per (model, dataset) pair; the higher of the two best validation accuracies is reported in Table 1.

The column abbreviations used in Table 1 are C100=CIFAR-100, C10=CIFAR-10, ChXR=Chest X-ray, MNIST=MNIST, Cars=cars196, DTD=dtd, STL10=stl10, INet-sk=imagenet-sketch, Cal101=caltech101, INet-v2=imagenet-v2, Scene=intel_scenes, SUN397=sun397, Flwr=oxford_flowers102, KITTI=kitti, ESAT=eurosat, CUB=cub200, PCam=patch_camelyon, Pet=oxford_iiit_pet, Food=food101, and INet=ImageNet (the pre-training reference).

A note on absolute baselines. Strict linear probing of frozen ImageNet-pretrained ResNet-34 features on small-image datasets such as CIFAR-100 (32×32 native resolution upsampled to 224×224) is known to yield substantially lower numbers than (i) full backbone fine-tuning, (ii) larger backbones, or (iii) dedicated low-resolution recipes. The baseline accuracies in Table 1 (e.g., 55.8% on CIFAR-100, 78.0% on CIFAR-10) are therefore consistent with the chosen protocol, which deliberately freezes the backbone to measure the transferability of the pretrained representation rather than the joint fine-tuning capacity. We use the same protocol for all compared models, so the relative differences (Δ Score in Table 1) remain meaningful even when the absolute baseline numbers are protocol-bound.

A.6 Compute and Runtime

All training in this work was conducted on a single server equipped with $8\times$ NVIDIA A100 GPUs. The software stack was Python 3.10, PyTorch 2.1.0, and CUDA 12.2.

Reproducing the models of this paper takes 9 variants over 11 runs, about 11 days of wall-clock time on this hardware: five distillation runs at roughly 1.5 days each (ED_Student, ED_SoftLabelOnly, and the Self / Gray / B&W Distilled variants) and six standard runs at roughly 0.6 days each (baseline, the Gray and B&W teachers, Adversarial Training, Noise Training, and SIN). The event teacher requires no training, since we use the public DiST checkpoint [5]. Verifying the central claim needs only two of these runs, the baseline and ED_Student, or about two days. To make cross-checking

cheaper we release the code, all trained checkpoints so that every evaluation in this paper can be reproduced without retraining, and an ImageNet-100 configuration for lightweight replication.

A.7 Summary of Trained Models and Clean Validation Accuracy

Table 10 consolidates, in one place, the clean top-1 validation accuracy of the teacher and student models used in the main comparisons. The DiST column reports each teacher's top-1 accuracy on the N-ImageNet validation split (event domain); the remaining three columns report the corresponding student's top-1 accuracy on the ImageNet validation split (RGB domain).

Two points are worth noting:

- The DiST ResNet-34 row (48.43%) corresponds to the publicly released checkpoint of [5] that we reuse as-is (Section A.2.1); the remaining DiST rows are teachers we trained in-house following the recipe of Section A.2.3.
- ED_SoftLabelOnly exhibits a wide spread (24.14–70.97%) because its student receives only the teacher's soft labels ($\alpha = 1$). When the teacher is strong relative to the student backbone (ConvNeXt-Base, RepLKNet31B), the signal transfers effectively; a weaker teacher paired with a weaker backbone (VGG16-BN) instead collapses the student. ED_Student improves on ED_SoftLabelOnly for every architecture but stays 3.10–13.11 percentage points below the corresponding baseline, consistent with the clean-accuracy / robustness trade-off reported throughout the main paper.

A.7.1 Teacher Confidence and the Ablation Confounder

The Event teacher (DiST ResNet-34) attains 48.43% top-1 on N-ImageNet, which is substantially lower than the ImageNet-1k accuracy of an RGB-trained ResNet-34 ($\approx 73.9\%$). One could therefore worry that the differing soft-label entropy of the teachers used in the ablation study (Section 6.1) confounds the modality comparison: the ED_Student behavior might reflect high-entropy KL targets rather than event-modality content. We argue this confound is bounded for three reasons.

(A) The control teachers are not uniformly highly confident. The Self Distilled teacher receives standard RGB inputs and converges close to baseline accuracy, but the Gray Distilled and B&W Distilled teachers are fed luminance-only and binarized inputs respectively at distillation time; these inputs carry strictly less information than the original RGB, so the per-input confidence of these control teachers is expected to be lower than that of the Self Distilled teacher. The asymmetry is therefore not a strict low-entropy-controls vs. high-entropy-DiST dichotomy.

(B) Entropy alone does not produce directional behavioral signatures. High-entropy soft labels, e.g., from temperature-scaled distillation [4] or label smoothing, can affect calibration and slightly aid generalization, but they are not known to selectively produce edge-tuned first-layer filters, cue-conflict shape bias, JPEG / pixelate trade-offs, or transfer gains on shape-relevant downstream tasks. The directional pattern observed for ED_Student matches the edge-based shape feature hypothesis, not a generic high-entropy regularization effect.

(C) ED_SoftLabelOnly is difficult to reconcile with the entropy-driven explanation. If the high-entropy DiST KL targets diluted the supervision, then ED_SoftLabelOnly ($\alpha = 1.0$, trained entirely via the high-entropy KL) should show weaker event-domain signatures than ED_Student. Empirically, ED_SoftLabelOnly exhibits the strongest signatures (highest filter correlation in Table 15, most pronounced shape bias in Appendix C.1), the opposite of what an entropy-confound explanation would predict.

A.8 Additional Non-Event Controls

Section 6.4 uses two further non-event supervisions. Both keep the architecture and the distillation recipe of ED_Student fixed, so that the teacher's input modality is the only experimental variable.

Canny_Distilled (RGB-derived edge teacher). Each RGB ImageNet image is converted to a binary edge map with `cv2.Canny(threshold1=` 100, `threshold2=` 200) and the map is replicated to three channels so that the input shape matches RGB. The teacher is a standard torchvision ResNet-34 trained from scratch on these edge maps; apart from the Canny thresholds, its recipe is

	Teacher	Student (ImageNet)		
Architecture	DiST	ED_SoftLabelOnly	ED_Student	baseline
ResNet-34	48.43	53.50	69.70	73.90
VGG16-BN	42.13	24.14	55.90	69.01
ConvNeXt-Base	52.42	70.97	78.40	81.50
ResNet-50	51.11	59.68	67.47	78.16
RepLKNet31B	51.61	67.43	76.61	83.45

Table 10: Clean top-1 accuracy (%) of the teacher and student models used in the main comparisons. The Teacher column (DiST) is evaluated on N-ImageNet validation (event domain); the Student columns are on ImageNet validation (RGB domain). The two additional non-event controls are listed separately in Table 11.

identical to the comparison teachers of Section A.2.2 (AdamW, bf16-mixed, 120 epochs, effective batch 512). We chose Canny because `cv2.Canny` is computed from Sobel gradient magnitudes followed by non-maximum suppression and hysteresis thresholding, so it stands in for the RGB-derived edge and gradient family (Canny, Sobel, Laplacian, gradient magnitude). The student is an RGB ResNet-34 distilled from this frozen teacher with exactly the ED_Student setting ($\alpha = 0.2$, $T = 3.0$, 120 epochs).

Synth_Distilled (augmentation-difference event proxy). The teacher input is a two-channel ON/OFF map computed from a single still RGB image:

1. Draw `RandomResizedCrop`(224, scale 0.5–1.0) and `RandomHorizontalFlip` once and share the result between the two views.
2. Apply `RandomAffine(translate=` $\pm 3\%$, i.e. ± 7 px at 224 px) and `ColorJitter`(brightness 0.15, contrast 0.15, saturation 0.10, hue 0.05) to each view with independent random draws.
3. Convert both views to grayscale (ITU-R 601) and take the difference $d = g_2 - g_1$.
4. Set ON $= \max(d, 0)$ and OFF $= \max(-d, 0)$, and stack them into a two-channel tensor normalized with mean $(0, 0)$ and standard deviation $(0.5, 0.5)$.

Sharing the crop is deliberate. When the crop is drawn independently for the two views, the difference degenerates into broad double-exposure blobs that do not follow object contours; sharing it leaves the ± 7 px translation and the mild photometric jitter as the only sources of difference, which yields sharp ON/OFF responses along contours. The teacher is a ResNet-34 whose `conv1` is replaced by a two-channel input layer (`out_channels`$= 64$, kernel 7, stride 2, padding 3), trained with the same optimizer, precision, batch size and schedule as above; the maps are generated on the fly inside the data loader. The student again follows the ED_Student setting. We stress that this proxy is derived from still RGB images: it is neither the output of an event simulator nor of a real event camera.

Control	Teacher input	Teacher	Student
Canny_Distilled	Canny edge map (3ch)	55.82	71.15
Synth_Distilled	augmentation-difference ON/OFF (2ch)	64.52	71.78

Table 11: Clean top-1 accuracy (%) of the two additional non-event controls. Teacher accuracy is measured on the correspondingly transformed ImageNet validation split; student accuracy is on the standard RGB validation split.

A.9 SOTA Training Recipe

To check that the reported gains are not an artifact of a weakly trained baseline (Section 6.4), we retrained both the baseline and ED_Student with a timm-style recipe: timm `resnet34` with `drop_path_rate`$= 0.05$ (stochastic depth), label smoothing 0.1, MixUp $\alpha = 0.2$ applied to

the cross-entropy branch only, RandAugment (`num_ops`$=2$, `magnitude`$=9$), RandomErasing $p = 0.25$, 300 epochs, AdamW with weight decay 0.05, 5 warmup epochs followed by cosine decay, effective batch size 1024, and bf16-mixed precision. The distillation setting of the SOTA ED_Student is unchanged from ED_Student ($\alpha = 0.2$, $T = 3.0$).

The learning rate is set to 75×10^{-5} for both models. This departs from the usual timm default, and we adopt it so that the SOTA baseline and the SOTA ED_Student differ only in the presence of the distillation term, which is what the comparison is meant to isolate.

	SOTA baseline	SOTA ED_Student
Clean top-1	75.92	70.80
ImageNet-Sketch top-1	26.98	28.94
Hue rotation, mean over 36–324°	63.96	65.76
Patch shuffle, grid $= 4$	53.82	37.48
Patch rotation, grid $= 4$	46.54	31.61
FGSM, mean over $\epsilon \in \{0.25, 0.5, 1\}/255$	34.74	28.70
PGD-10, mean over the same ϵ	16.65	16.30
Bandpass accuracy drop at 0.95π (pp, lower is better)	25.27	21.07
Linear probe, mean over 19 tasks	74.19	75.45

Table 12: Baseline and ED_Student retrained under the SOTA recipe. ED's advantage persists on shape, color, high-frequency-band robustness and downstream transfer, but not on the adversarial axes, where the stronger augmentation already makes the SOTA baseline robust at these perturbation sizes.

A.10 Cross-Method Comparison on the Figure-1 Axes

Section 6.4 uses one metric per axis for compactness. Table 13 expands that comparison onto the axes summarized in the right panel of Figure 1, reporting two color axes (hue rotation and grayscale) and two shape axes (cue-conflict and ImageNet-Sketch), because each pair probes the same property from a different direction. Synth_Distilled is compared on the four axes of Table 2.

	LP mean	Hue Avg.	Grayscale	Cue-conflict	Sketch	Bandpass @0.95π	Adv. AUC
baseline	73.5	61.8	62.4	0.212	23.7	44.3	6.2
ED_Student	**75.1**	65.4	64.6	0.375	28.1	57.5	8.9
ED_SoftLabelOnly	—	52.4	52.6	0.487	24.5	45.8	9.1
Canny_Distilled	73.7	**66.1**	**65.1**	0.270	27.9	35.6	7.5
Adversarial Training	71.7	50.6	49.5	**0.572**	21.3	**64.5**	**24.1**
Noise Training	72.7	62.6	63.3	0.258	25.4	58.2	11.5
SIN	72.6	62.3	62.5	0.428	**30.8**	61.8	6.5

Table 13: Comparison on the axes of Figure 1. LP mean: linear-probe accuracy averaged over the 19 tasks of Table 1; the linear probe is run for the five models compared in that table plus Canny_Distilled. Hue Avg.: top-1 averaged over ten hue-rotation angles including $0°$. Grayscale: top-1 on the grayscale validation split. Cue-conflict: shape bias under the 16-class restricted-argmax protocol. Sketch: ImageNet-Sketch top-1. Bandpass: top-1 under band-limited noise centered at 0.95π (higher is better). Adv. AUC: area under the PGD accuracy curve over $\epsilon \in \{0, 0.25, 0.5, 1\}/255$, scaled by 100. Bold marks the best value in each column.

No single supervision dominates. Adversarial Training is the most robust on both the frequency and the adversarial axes but loses 11.2 pp on hue rotation, 12.9 pp on grayscale and 1.8 pp on downstream transfer relative to the baseline. SIN attains the highest top-1 accuracy on ImageNet-Sketch but does not improve downstream transfer. Canny_Distilled matches ED_Student on the color and sketch axes, yet its cue-conflict shape bias (0.270) stays much closer to the baseline (0.212) than to ED_Student (0.375), and it is the only model whose bandpass accuracy falls below the baseline. ED_Student is the only supervision that improves over the baseline on color, shape, high-frequency-band robustness, and downstream transfer simultaneously.



Figure 6: Visual examples of hue-rotated inputs. A single ImageNet validation image is shown after rotating its hue from $0°$ to $330°$ in $30°$ steps (12 images). The label H denotes the relative hue value in HSV space ($H \in [0, 1)$, with 1 corresponding to $360°$).

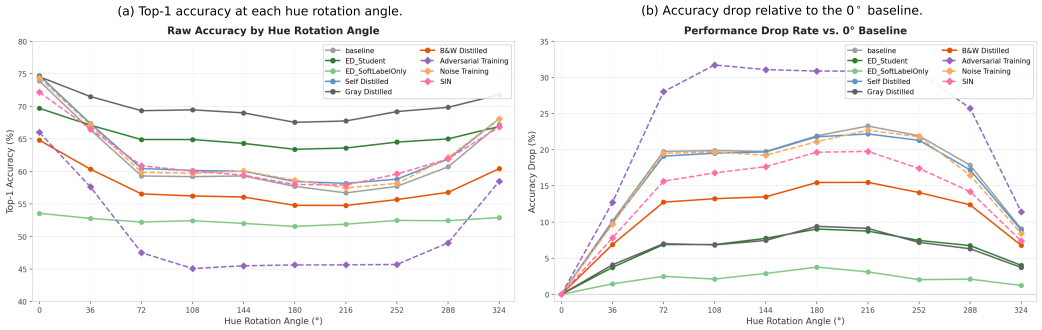


Figure 7: Hue-rotation robustness across all 9 models. a) Top-1 accuracy at each hue angle ($0°$–$324°$ in $36°$ steps). b) Accuracy drop relative to $0°$.

B Extended Analysis of Color Invariance

In this appendix, we present additional quantitative and qualitative analyses of the Color and Hue Invariance discussed in Section 4.1 of the main paper, which did not fit on the main pages. Specifically, we provide three stages of validation: (i) a comparison of hue-rotation robustness across the nine models of Figure 7 (Appendix B.1), (ii) behavior under grayscale conditions in which only color information is removed (Appendix B.2), and (iii) an internal-representation analysis via visualization of early-layer feature maps under hue rotation (Appendix B.3). Behavior under B&W, Canny, and silhouette conditions, in which texture is further removed beyond grayscale and only shape information remains, is consolidated in Appendix C, organized from the perspective of shape bias.

B.1 Across-Model Comparison under Hue Rotation

Figure 7 shows, for all 9 models trained in this work, (a) raw top-1 accuracy at each hue angle and (b) the accuracy drop relative to the $0°$ baseline, when the hue of the entire input image is rotated from $0°$ to $324°$ in $36°$ steps. Figure 2 (a) of the main paper shows only 3 of the models (baseline, ED_Student, ED_SoftLabelOnly), so here we summarize the behavior of the remaining 6 models (Self Distilled, Gray Distilled, B&W Distilled, Adversarial Training, Noise Training, SIN).

Adversarial Training shows the largest accuracy drop among all models. The perturbations PGD training assumes are small per-pixel high-frequency noises in the L_∞ norm, whereas hue rotation is a structured global transformation across the entire color space and is therefore outside the scope of the attack model. Moreover, by suppressing reliance on high-frequency texture cues, adversarial training can, as a side effect, strengthen reliance on the remaining color-related cues.

Self Distilled, Noise Training, and the SIN model all show only roughly the same hue-rotation robustness as the baseline. Notably, the SIN model shows little color invariance despite acquiring a strong shape bias via Stylized-ImageNet, indicating that acquiring shape bias does not necessarily entail color invariance.

B&W Distilled and Gray Distilled, in which color information has been heavily removed on the teacher side in advance, show stronger color invariance than the baseline, but their magnitude remains comparable to or less than that of ED_Student. This supports that the color invariance acquired by the Event-distilled model cannot be fully explained by inheritance of an inductive bias



Figure 8: Visual examples of each transformation. a) RGB, b) grayscale, c) sketch, d) B&W (Otsu thresholding), e) Canny, f) silhouette, g) synthetic event. The last panel is the two-channel ON/OFF map used to train the Synth_Distilled teacher (Section A.8), rendered with ON in red and OFF in blue; it is the grayscale difference between two augmented views of a) and is not the output of an event simulator or of a real event camera.

from removing color information from the teacher alone, and is rooted in dependence on brightness gradients (edges) intrinsic to the event domain (see Section 6.1 of the main paper).

B.2 Robustness under Grayscale Inputs

	ImageNet-val	grayscale	sketch	B&W	canny	silhouette
baseline	73.90	62.45	23.70	41.39	10.83	3.30
ED_Student	69.70	64.64	28.08	49.01	13.27	**5.43**
ED_SoftLabelOnly	53.50	52.58	24.47	34.87	11.45	5.42
Self Distilled	**74.72**	62.97	24.43	42.69	11.59	3.24
Gray Distilled	74.60	**69.51**	25.60	45.56	7.87	3.63
B&W Distilled	64.81	56.20	23.47	45.54	12.12	3.22
Adversarial Training	66.02	49.50	21.27	29.78	4.43	3.28
Noise Training	74.35	63.32	25.39	45.39	10.28	4.09
SIN	72.18	62.53	**30.84**	**51.62**	**25.77**	5.38

Table 14: Top-1 accuracy (%) on the ImageNet validation set with color/texture information progressively stripped. Bold marks the highest in each column, underline the second. Referenced from both Appendix B.2 and C.2. Visual examples in Figure 8.

Table 14 shows top-1 accuracy for all 9 models under conditions in which the ImageNet validation set is preprocessed with five transformations (grayscale, sketch, B&W, Canny, silhouette). Note that, unlike the others, the sketch column is not a transformation derived from ImageNet-1k but the accuracy on the validation split of the independent ImageNet-Sketch dataset [85]. In this subsection, we focus on the grayscale column, which removes only color information (the behavior under shape-only conditions following binarization is treated in Appendix C.2). Visual examples of each transformation are shown in Figure 8.

Under grayscale, Gray Distilled, in which color information has been removed on the teacher side in advance, attains the highest accuracy, followed by ED_Student. Compared to baseline, the Event-distilled model shows a clear advantage, evidencing an inference path that does not require color information. The SIN model is at the baseline level, confirming that acquiring shape bias by another route does not in itself confer color invariance. By contrast, Adversarial Training degrades to roughly 13 pp below baseline already at the grayscale stage, indicating that the feature-representation distortion accompanying adversarial training becomes apparent under the loss of color information.

B.3 Layer-wise Feature Maps under Hue Rotation

To directly confirm at which stage of the representation space the color invariance emerges, we visualized the feature maps in the early convolutional layers (Layer 0–2) of ResNet-34 under hue-rotated inputs. Figure 9 shows the baseline model and Figure 10 the ED_SoftLabelOnly model: each figure displays the responses of every layer and channel as the hue of the input image is rotated from 0° to 330° in 30° steps. The red boxes in each figure highlight channels whose response strength or spatial distribution varies substantially with the input hue: channels without color invariance.

In the baseline model, many channels already at Layer 0 are flagged in red, indicating responses strongly dependent on input color information. In contrast, in the ED_SoftLabelOnly model, the early-layer feature maps remain almost invariant under the same hue-rotated inputs, and only a very

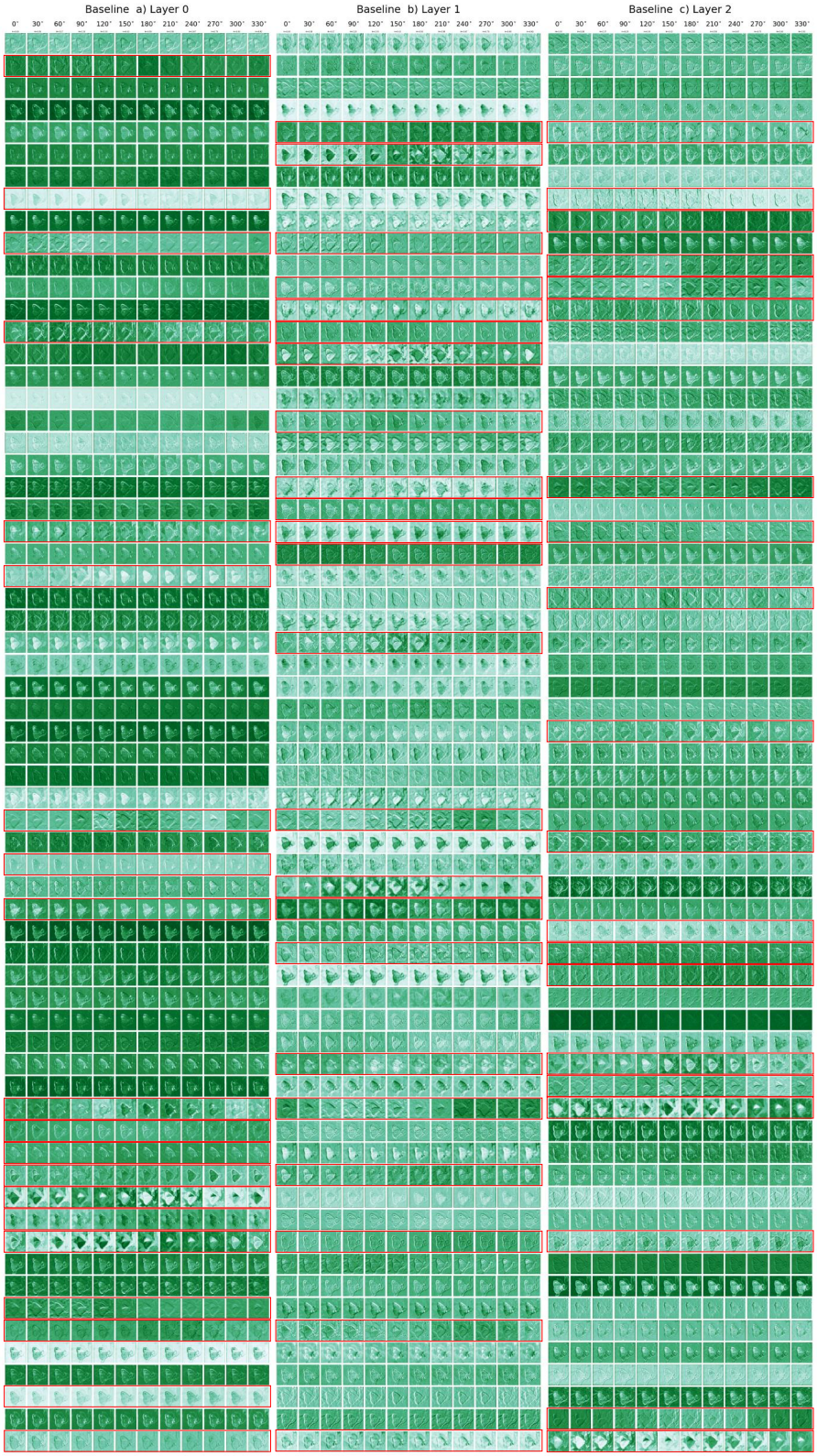


Figure 9: Feature maps in the early layers (Layer 0–2) of the baseline model. Each column corresponds to a hue rotation angle (0°–330° in 30° steps), and each row corresponds to one channel. Red boxes mark channels whose response varies substantially with hue change (not color-invariant). See Figure 6 for visual examples of the hue-rotated inputs.

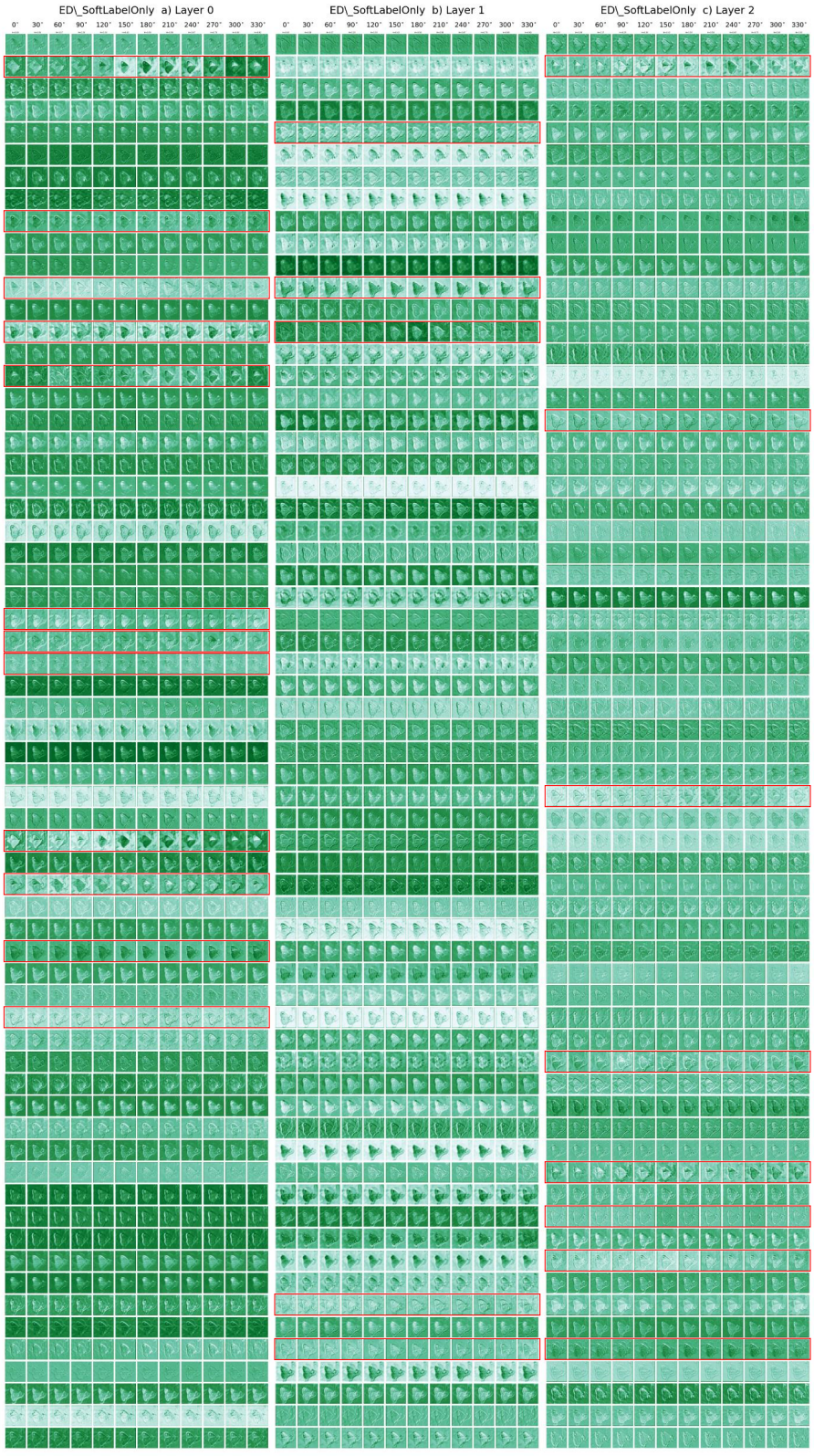


Figure 10: Early-layer feature maps of ED_SoftLabelOnly. Same format as Figure 9. Responses are notably invariant to hue rotation; only a few channels are flagged in red.

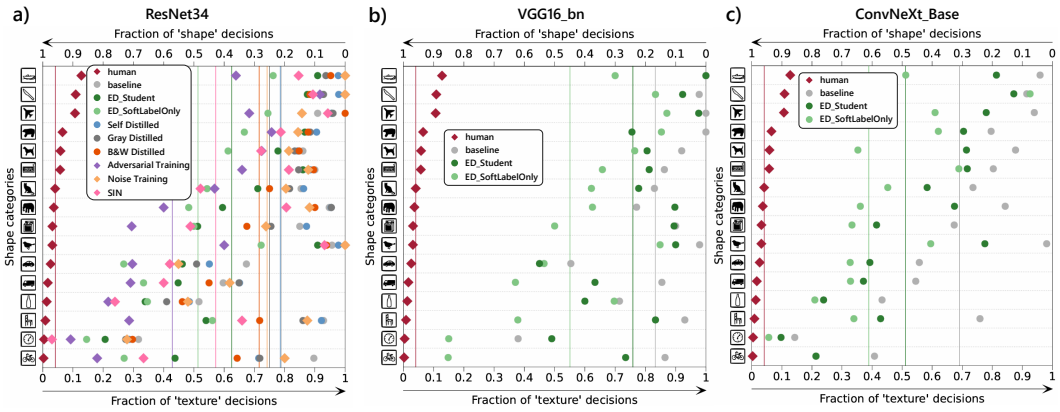


Figure 11: Cue-conflict experiments across architectures and models. a) Shape-bias matrix-plot of all 9 ResNet-34 models (baseline, ED_Student, ED_SoftLabelOnly, Self/Gray/B&W Distilled, Adversarial Training, Noise Training, SIN) and the human reference, computed using the bethgelab/model-vs-human 16-class restricted-argmax protocol [1]. b) VGG-16 and c) ConvNeXt-Base, each with the three models baseline / ED_Student / ED_SoftLabelOnly. Each panel plots the texture-versus-shape decision ratio per category as a scatter plot.

small number of channels are flagged in red. In other words, this model is unaffected by changes in input color and consistently extracts only geometric structure (edges) based on the Structural Prior discussed in Section 5.1 of the main paper. This is confirmed at the layer level via visualization. The result demonstrates, at the per-pixel response level, that the color invariance of the Event-distilled model arises not from post-processing at the output layer but from a representation transformation at the stage closest to the input.

C Extended Analysis of Shape Bias

In this appendix, we provide additional quantitative and qualitative analyses of the High Shape Bias discussed in Section 4.2 of the main paper, which did not fit on the main pages. Specifically, we examine the topic from four perspectives: (i) cross-architecture and cross-model comparison of cue-conflict experiments (Appendix C.1), (ii) accuracy comparison under shape-only conditions in which color and texture information have been stripped to the extreme (Appendix C.2), (iii) behavior when the macroscopic shape is destroyed by patch-shuffle / patch-rotation (Appendix C.3), and (iv) internal-representation analysis through a functional taxonomy of first-layer convolutional filters (Appendix C.4).

C.1 Cue-Conflict: Across Architectures and Models

We extend the cue-conflict result on ResNet-34 shown in Figure 2 (b) of the main paper along two axes: (i) across architectures (VGG-16, ConvNeXt-Base) and (ii) across models within ResNet-34 (all 9 models).

Across architectures (Figure 11b, c). For both VGG-16 and ConvNeXt-Base, the Event-distilled models (ED_Student, ED_SoftLabelOnly) shift clearly toward the shape side (left) of the plot relative to baseline, matching the trend on ResNet-34 shown in Figure 2 (b) of the main paper. This result constitutes one piece of evidence for the architectural generality discussed in Section 6.2 of the main paper.

Across models on ResNet-34 (Figure 11a). Under the official 16-class restricted-argmax protocol, the overall shape bias is ordered as Adversarial Training (0.572) > ED_SoftLabelOnly (0.487) > SIN (0.428) > ED_Student (0.375) > B&W Distilled (0.284) > Noise Training (0.258) > Gray Distilled (0.251) > Self Distilled (0.214) > baseline (0.212), with the human reference at 0.959. SIN sits between ED_Student and ED_SoftLabelOnly, so the strength of its shape bias is comparable to that of the Event-distilled models. However, the Event-distilled model differs from both Adversarial Training and SIN in its side-effect profile: Adversarial Training exhibits the largest accuracy drop under hue rotation among all models (Appendix B.1) and falls below the baseline on 12 of 19 downstream transfer tasks (Table 1); SIN's hue-rotation robustness is at the baseline level (Appendix B.1)

and 12 of 19 downstream tasks fall below baseline, while it achieves the highest accuracy on the shape-only inputs (B&W / Canny / sketch columns of Table 14), even surpassing ED_Student. Unlike these two routes, the Event-distilled model preserves color invariance (Appendix B) and transfer performance (Table 1) while still acquiring a strong shape bias, and therefore offers an advantage when both the magnitude of the shape bias and the trade-off across the evaluated metrics are taken into account.

C.2 Robustness under Shape-Only Inputs

Of Table 14 and Figure 8 shown in Appendix B.2, this subsection focuses in particular on the four conditions in which most texture and color information have been stripped and only shape information remains (sketch, B&W, Canny, silhouette).

Under these four conditions, the SIN model, whose shape learning is directly induced by Stylized-ImageNet co-training, attains the highest accuracy on three of them (sketch, B&W, Canny), with ED_Student following as second on each. On silhouette, ED_Student is highest, followed essentially in a tie by ED_SoftLabelOnly and SIN. Even though the Event-distilled model performs no explicit shape learning, its ability to infer class from shape information alone reaches a level comparable to SIN, and clearly exceeds Self Distilled and Noise Training, both of which use RGB training inputs under all conditions. Including the remaining control models that did not fit in the main paper, this corroborates the claim of Section 5.3 of the main paper that the Event-distilled model can perform stable class inference from object shape information alone, without requiring superficial cues such as color or texture.

In contrast, Adversarial Training falls below baseline on all four shape-extraction conditions. The fact that adversarial training, despite having the effect of suppressing high-frequency texture dependence, does not lead to inference from shape information suggests that the essence of robustness lies not only in texture suppression but in selective acquisition of shape features.

C.2.1 Generation of Canny and Silhouette Evaluation Inputs

The Canny and silhouette columns of Table 14 are evaluated on offline-generated parallel trees of the ImageNet validation split (one image per source image, saved as PNG to avoid JPEG-compression artifacts). The B&W column is generated by the Otsu thresholding pipeline already documented in Section A.4.3.

Canny. Each ImageNet validation image is converted to grayscale (`PIL.Image.convert("L")`) and normalized to float32 in $[0, 1]$. A Gaussian blur with $\sigma = 1.0$ and kernel size $\lceil 6\sigma + 1 \rceil = 7$ (forced odd) is applied, after which the image is rescaled to `uint8` and passed to `cv2.Canny` with low/high thresholds $0.35 \times 255 \approx 89$ and $0.9 \times 255 \approx 230$. The resulting binary edge map is saved as PNG.

Silhouette. Each image is loaded as RGB and passed through the `rembg` library with the `u2net` session for foreground extraction. The returned RGBA tensor's alpha channel is binarized by the rule $\text{alpha} > 10 \mapsto 255$, $\text{alpha} \leq 10 \mapsto 0$, yielding a two-valued mask that is saved as PNG.

Both pipelines preserve the strict 0/255 binary structure of the saved images, allowing the validation loader to feed them directly into the same evaluation transform (`Resize`(256) + `CenterCrop`(224) + ImageNet `Normalize`) used for the RGB / grayscale columns.

C.3 Patch-Shuffle and Patch-Rotation Experiments

As further evidence of shape bias, we conducted experiments that destroy the continuity of macroscopic edges by dividing the image into a grid and shuffling / rotating each patch. Patch shuffling and rotation are widely adopted diagnostic procedures: they preserve the local texture inside each patch while selectively destroying the global shape information that spans across patches, allowing one to disentangle whether a model relies on within-patch local cues or on across-patch macroscopic shape [72]. By varying the grid granularity, we quantitatively evaluate the spatial scale of the cues on which the Event-distilled model relies. Figure 12 shows the top-1 accuracy progression as the patch size (grid granularity) is varied.

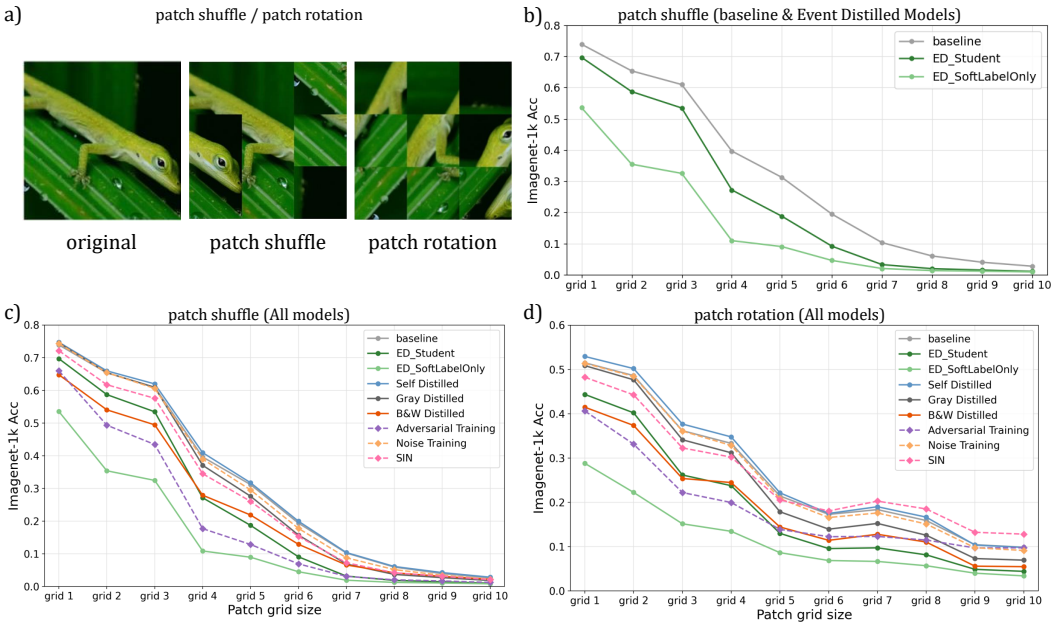


Figure 12: Patch-shuffle / patch-rotation experiments. a) Visual examples. b) Top-1 accuracy under patch shuffle as a function of grid granularity (baseline vs. Event-distilled models). c) Accuracy under patch shuffle for all 9 models. d) Same for patch rotation.

When the grid is coarse, the local texture inside each patch is preserved, so the texture-dependent baseline maintains relatively high accuracy. However, as the grid becomes finer and the continuity of macroscopic edges is broken, the Event-distilled model's accuracy drops more sharply than the baseline's. This is direct evidence that the information source on which the Event-distilled model relies for inference is not local texture within a patch but the macroscopic shape structure that holds across patches. The same trend is observed under patch rotation, and in both conditions the Event-distilled model exhibits steeper accuracy degradation than even Gray Distilled and B&W Distilled, which are distilled from teachers stripped of color information. This confirms that the Event-distillation-specific dependence on macroscopic shape continuity is a property distinct from the other distillation variants. The SIN model, which acquires shape bias through Stylized-ImageNet, shows accuracy curves comparable to the baseline under both patch-shuffle and patch-rotation conditions, indicating that even when shape bias is acquired through an alternative route, sensitive dependence on macroscopic shape continuity does not necessarily follow.

As part of the architectural generality validation discussed in Section 6.2 of the main paper, we also conduct the same patch-shuffle / patch-rotation experiments on VGG-16, ConvNeXt-Base, ResNet-50, and RepLKNet31B. As shown in Figure 13, across all architectures the Event-distilled model exhibits the same characteristic behavior as on ResNet-34: its accuracy drops more sharply than baseline as the patch size becomes finer. This confirms that the Event-distilled model's dependence on macroscopic shape continuity arises consistently regardless of the architectural design.

C.4 Functional Taxonomy of First-Layer Filters

In this subsection, we provide a more detailed analysis of the functional taxonomy of first-layer convolutional filters introduced in Section 5.1 of the main paper. While Figure 3 (a) of the main paper shows only a few excerpted channels due to space constraints, Figure 14 visualizes all 64 channels of the first convolutional layer of each model. The Event-distilled model (especially ED_SoftLabelOnly) is observed to dominantly acquire smooth, spatially highly correlated structures across the entire filter layer, not just an excerpt.

The acquired 64 filters can be broadly classified into the following three types based on their spatial and color response characteristics.

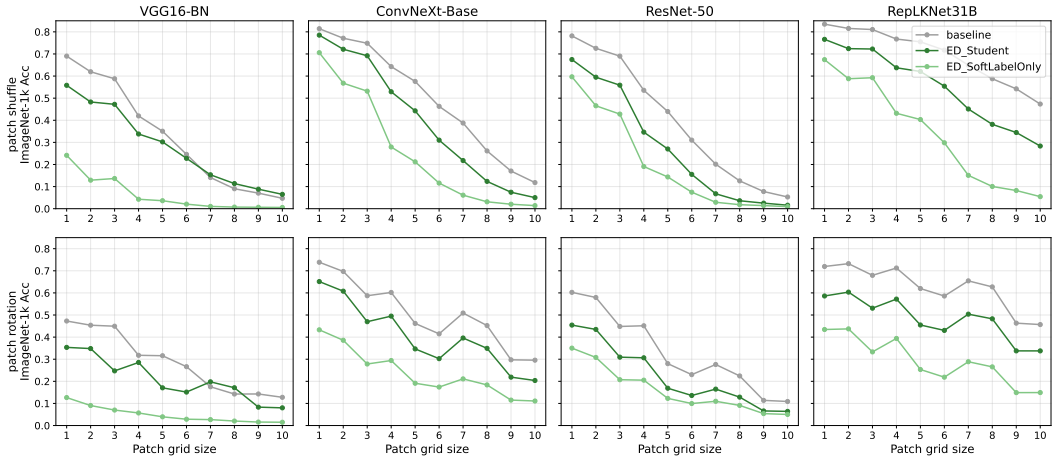


Figure 13: Cross-architecture comparison of patch-shuffle / patch-rotation experiments. For VGG-16, ConvNeXt-Base, ResNet-50, and RepLKNet31B, we report the top-1 accuracy progression of baseline / ED_Student / ED_SoftLabelOnly as the number of patch divisions varies. Top: patch shuffle. Bottom: patch rotation.

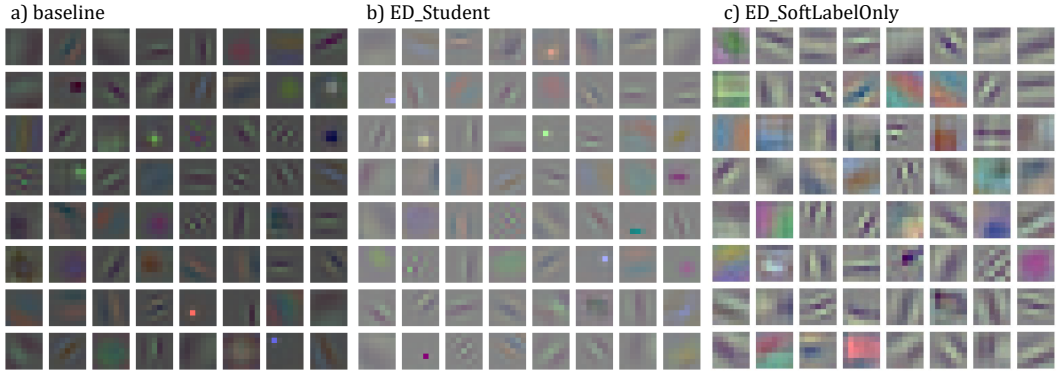


Figure 14: All 64 channels of the first convolutional layer of ResNet-34. a) baseline, b) ED_Student, c) ED_SoftLabelOnly. Whereas Figure 3 (a) shows excerpted channels, this figure shows the layer-wide tendency.

- Monochromatic Global Filters (e.g., 1st column of Figure 3 (a)): These form smooth, monochromatic gradients that span the entire receptive field, and serve to smooth out local color variation, thereby estimating broad illumination conditions or macroscopic color distribution of the background. They are believed to contribute to adaptive normalization that highlights structural luminance changes in subsequent layers.
- Color-Opponent Structural Filters (e.g., 2nd column of Figure 3 (a)): These have a structure smoothly partitioned by two near-complementary colors, exhibiting response characteristics analogous to color-opponent responses in the human visual system. They suppress fine textures while extracting macroscopic boundaries between objects with different color characteristics.
- Gabor-like Edge Detectors (e.g., 3rd–4th columns of Figure 3 (a)): These have orientations and spatial-frequency characteristics similar to Gabor filters, retaining sensitivity to the steep structural gradients (edges) that constitute shape, while suppressing unstructured high-frequency texture. They faithfully reflect the brightness gradients intrinsic to the event sensor and directly extract macroscopic object contours (edge-based shape feature).

For quantitative support, the filters' channel-wise correlation (RGB Correlation) and spatial autocorrelation (Spatial Autocorrelation) are shown in Table 15. ED_SoftLabelOnly's filters have substantially higher channel correlation and spatial autocorrelation than baseline (0.747 vs. 0.549 for RGB Correlation; 0.575 vs. 0.391 for Spatial Autocorrelation). ED_Student's first-layer correlations are close to baseline (0.544 / 0.398); we discuss this dichotomy in Appendix C.4.1, where we argue that

the strong CE branch dominates first-layer shaping for ED_Student while the underlying mechanism is distributed across the representation rather than localized in the first layer. The high correlation of ED_SoftLabelOnly indicates (i) that all RGB channels synchronously extract the same geometric structure, and (ii) that the model has acquired a smooth spatial structure that emphasizes continuity between adjacent pixels, quantitatively reinforcing the structural prior discussed in the main paper.

Correlation Type	RGB Correlation	Spatial Autocorrelation
baseline	0.549	0.391
ED_Student	0.544	0.398
ED_SoftLabelOnly	0.747	0.575

Table 15: Correlation values for the 64 channels of the first convolutional layer of ResNet-34. RGB Correlation is the cross-channel correlation, and Spatial Autocorrelation is between adjacent pixels. Higher values indicate smoother, more isotropic filter structures.

C.4.1 Robustness of the Mechanistic Claim to the Augmentation Pipeline

Since ED_SoftLabelOnly is trained entirely via the KD branch (the weak pipeline of Table 8), one might wonder whether the absence of RandAugment and RandomErasing alone, rather than event distillation, is what produces the smoother filters and higher correlation values reported in Section 5.1 of the main paper. We argue that this confound is bounded for three reasons.

(A) The augmentation pipeline is a perturbation modulator, not a smoothing driver. A weak pipeline does not actively push filters toward smoothness; it merely omits perturbations. The gradient signal that organizes the filters comes from the loss target, here the event teacher's KL divergence, not from the augmentation itself. RandAugment [93] and RandomErasing [94] are designed to broaden the input distribution rather than to bias filters toward high-frequency texture, so their removal is unlikely on its own to account for the smooth, oriented Gabor-like structures observed in Figure 14.

(B) Filter shape is determined by the loss target and the input statistics, not by the augmentation strength. Classical work in unsupervised learning of natural image statistics shows that Gabor-like filters emerge from edge-rich input distributions under a sparse-coding objective [95]. CNN first-layer visualizations across diverse training recipes, from AlexNet [96] to deconvolutional analyses [97], consistently show that texture-tuned first-layer filters arise under standard CE training on natural images, regardless of whether strong augmentation is applied. The Gabor-organized, edge-oriented structure of ED_SoftLabelOnly therefore reflects the upstream input statistics of N-ImageNet (a flatter PSD and edge-richer distribution; Appendix E) propagated through the KL distillation, not the absence of augmentation. This is consistent with prior work on the origins of texture bias [1, 53], which identifies the loss / data interaction, not augmentation noise, as the primary driver of low-level representation shape.

(C) The mechanistic claim is not localized in first-layer scalar correlation. ED_Student receives 80% strong-pipeline CE and only 20% KL; its first-layer scalar correlations (RGB 0.544, Spatial 0.398) are accordingly close to baseline (0.549 / 0.391), which is consistent with the strong CE branch dominating first-layer shaping; we did not measure the relative gradient contributions of the two terms. Yet ED_Student still exhibits the full set of behavioral signatures attributed to the structural prior: color invariance (Section 4.1), shape bias (Section 4.2), high-frequency-noise robustness (Section 4.3), the JPEG / pixelate trade-off (Section 5.2), shape-only inference (Section 5.3), and downstream transfer (Section 6.3). The mechanism is therefore a representation shift that is not localized in first-layer smoothness; the filter analysis serves as one diagnostic among several converging pieces of evidence.

C.4.2 Where the Filter Change Is Localized

Figure 3 (a) shows individual first-layer filters, whereas Figure 3 (b) aggregates scalar statistics over all of them, and the two do not obviously agree. A layer-by-layer measurement resolves the tension.

Per-layer alignment between students. For each of the 36 convolutional layers of the students we compute, for every filter of one model, the maximum cosine similarity to the filters of another model, and average. At `conv1` all pairs are high (0.53–0.58), reflecting the shared RGB

stem. At `layer1` the pair ED_Student–ED_SoftLabelOnly reaches 0.161, above every other pair (baseline–ED_Student 0.142, baseline–ED_SoftLabelOnly 0.135, baseline–Canny_Distilled 0.144, ED_Student–Canny_Distilled 0.141, ED_SoftLabelOnly–Canny_Distilled 0.136): the two event-supervised students are measurably closer to each other than to either the baseline or the edge-map control. The gap narrows at `layer2` and is indistinguishable at `layer3` and `layer4`.

The structural change induced by event supervision is therefore localized in the early layers, principally `layer1`. This layer-wise similarity is a separate diagnostic from the first-layer correlation statistics of Figure 3 (b), which aggregate over the filters of a single layer, and the two should not be read as measuring the same quantity.

D Extended Analysis of High-Frequency Robustness

In this appendix, we present additional analyses of Robustness to High-Frequency Perturbations discussed in Section 4.3 of the main paper, which did not fit on the main pages. Specifically, we examine four aspects: (i) hyperparameters used for the PGD evaluation (Appendix D.1), (ii) response to band-limited random noise (Appendix D.2), (iii) comprehensive evaluation against all 19 ImageNet-C corruptions (Appendix D.3), and (iv) interpretation of the Event-distilled model's vulnerability based on the frequency characteristics of additive noise (Appendix D.4).

D.1 PGD Evaluation Hyperparameters

The PGD adversarial-robustness curves shown in Figure 2 (c) of the main paper and in Figure 5 (Section 6.2) use a common evaluation pipeline, distinct from the PGD recipe used to train the Adversarial Training baseline (Appendix A.3.3, $\epsilon = 8/255$, $\alpha = 2/255$, $K = 10$).

- **Attack**: L_∞ PGD with cross-entropy loss and sign-of-gradient updates.
- **Iterations**: $K = 10$ steps.
- **Step size**: $\alpha = (2.5 \cdot \epsilon)/K$, a conventional choice that slightly overshoots so that the projection is hit at every step independently of ϵ.
- **Random restart**: a single restart, with the perturbation initialized uniformly in $[-\epsilon, +\epsilon]$ and clipped back to $[0, 1]$ before iteration 1.
- **Image space**: the attack operates in the $[0, 1]$ pixel space; ImageNet normalization is performed inside a model wrapper (`NormalizedModel`) so that the same attack budget ϵ corresponds to the same physical pixel perturbation across all compared models, irrespective of training-time normalization statistics.
- **Evaluation set**: the standard ImageNet validation split, processed by `Resize`(256) + `CenterCrop`(224) + `ToTensor` (no normalization outside the wrapper).
- ϵ **grid**: $\{0,\ 0.25/255,\ 0.5/255,\ 1/255\}$ in Figure 2 (c) and Figure 5 (the FGSM panels share the same ϵ grid). Larger budgets are not reported because all models collapse beyond $1/255$ in this single-restart, low-iteration setting.

The same hyperparameters are used for every model in the FGSM/PGD comparisons; only the model weights differ across runs.

D.2 Frequency-Band Random Noise

Following the procedure of Park and Kim [98], given an input image x_0, we generate a perturbed input x_{noise} that contains only a specified frequency band of noise as

$$x_{\text{noise}} = x_0 + \mathcal{F}^{-1}(\mathcal{F}(\delta) \odot M_f)\,, \tag{2}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ are the 2D Fourier transform and its inverse, δ is Gaussian noise, M_f is a band-pass mask that lets only the band centered at $f \in [0, \pi]$ pass, and $\odot$ denotes element-wise product. Applying x_{noise} to the ImageNet validation set and measuring the top-1 accuracy as f varies allows us to evaluate which frequency-band noise each model is fragile or robust to.

The concrete settings are $\epsilon = 0.2$ for the noise magnitude, an L_∞ band-pass mask (a square ring in the frequency plane), and a bandwidth of 0.1π; the eleven center frequencies are $f/\pi \in$

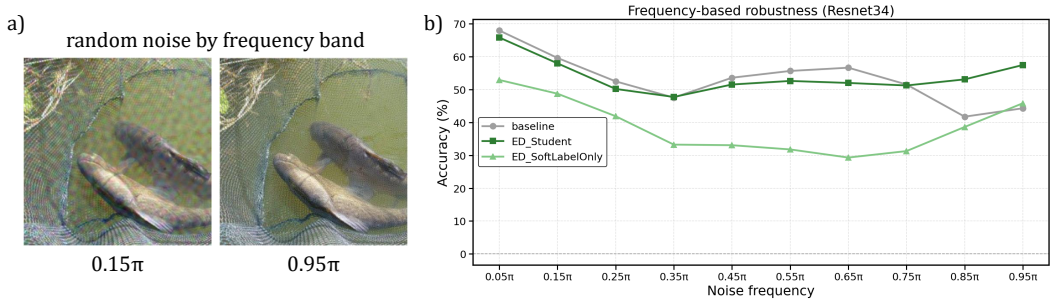


Figure 15: Robustness to frequency-band random noise. a) Examples of inputs perturbed by band-limited noise (*left*: center frequency 0.15π, *right*: 0.95π). b) Top-1 accuracy (absolute) under noise concentrated at each band; horizontal axis is the center frequency of the added noise (0–π).

$\{0, 0.05, 0.15, \ldots, 0.95\}$. These choices matter: replacing the L_∞ ring with an L_2 (circular) one and halving ϵ shrinks the measured drops by roughly two orders of magnitude, so band-pass numbers are only comparable within a fixed protocol.

Figure 15 (a) shows examples of noisy inputs at low (0.15π) and high (0.95π) center frequencies. The low-frequency noise gently modulates the global brightness of the image, whereas the high-frequency noise appears as fine granular patterns on the surface. Figure 15 (b) plots the absolute top-1 accuracy as the noise's center frequency is varied from 0π to π. In particular, in the high-frequency band (0.75π and above), the Event-distilled model degrades markedly less than the baseline: at 0.95π its accuracy drops by 12.2 pp against the baseline's 29.6 pp. Both models are still affected, so the difference is one of degree rather than of invariance. This is consistent with the smooth low-pass-filter-like response discussed in Section 5.1 of the main paper, indicating that the Event-distilled model does not rely on high-frequency texture as an inference cue. On the other hand, in the low- and mid-frequency bands (0π–0.5π), the gap between models narrows, indicating that the Event-distilled model's advantage is concentrated in the high-frequency range; in the low–mid range, both models are similarly affected because the Event-distilled model itself relies on these bands for edge extraction, as detailed in the noise frequency decomposition of Appendix D.4.

D.3 Comprehensive Evaluation on ImageNet-C

To comprehensively verify the spectral trade-off presented in Section 5.2 of the main paper, Figure 16 reports the accuracy evaluation against all 19 corruptions comprising ImageNet-C. In the figure, corruptions are visualized in four groups: Blur, Digital, Noise, and Weather.

We summarize the observations that reinforce the main paper's claims for each group. The clearest gain appears on perturbations that selectively remove high-frequency texture on the image surface while largely preserving the macroscopic contours of objects. On Digital corruption jpeg_compression and Weather corruption fog, the Event-distilled model consistently exceeds baseline, confirming that such distribution shifts are favorable to the Event-distilled model, which does not rely on texture. This supports the spectral-trade-off hypothesis of Section 5.2 in the main paper.

On the other hand, on Blur (gaussian_blur, defocus_blur, motion_blur, glass_blur, zoom_blur), the Event-distilled model's advantage is limited, and at high severities it can fall below baseline. Blur smooths not only high-frequency texture but the edges themselves, so for the Event-distilled model, which strongly depends on edges, blur attenuates its very inference cue. This observation indicates that the benefit of the Event-distilled model is restricted not to texture removal in general, but to texture removal under preserved edges, reflecting its strong dependence on the edge-based shape feature.

In contrast, on Noise (gaussian_noise, shot_noise, impulse_noise, speckle_noise), the Event-distilled model degrades to roughly the same accuracy as baseline or worse. Furthermore, on Digital corruptions that physically damage the structure itself (pixelate, contrast, elastic_transform), at low severities the Event-distilled model maintains higher accuracy than baseline, but in regions where severity increases and the original edge continuity is lost, a reversal occurs and it falls below baseline. These vulnerabilities support the claim of Section 5.2 of the main paper, that the Event-distilled model is strongly affected when the geometric structure and edge continuity on which it relies for inference are destroyed.

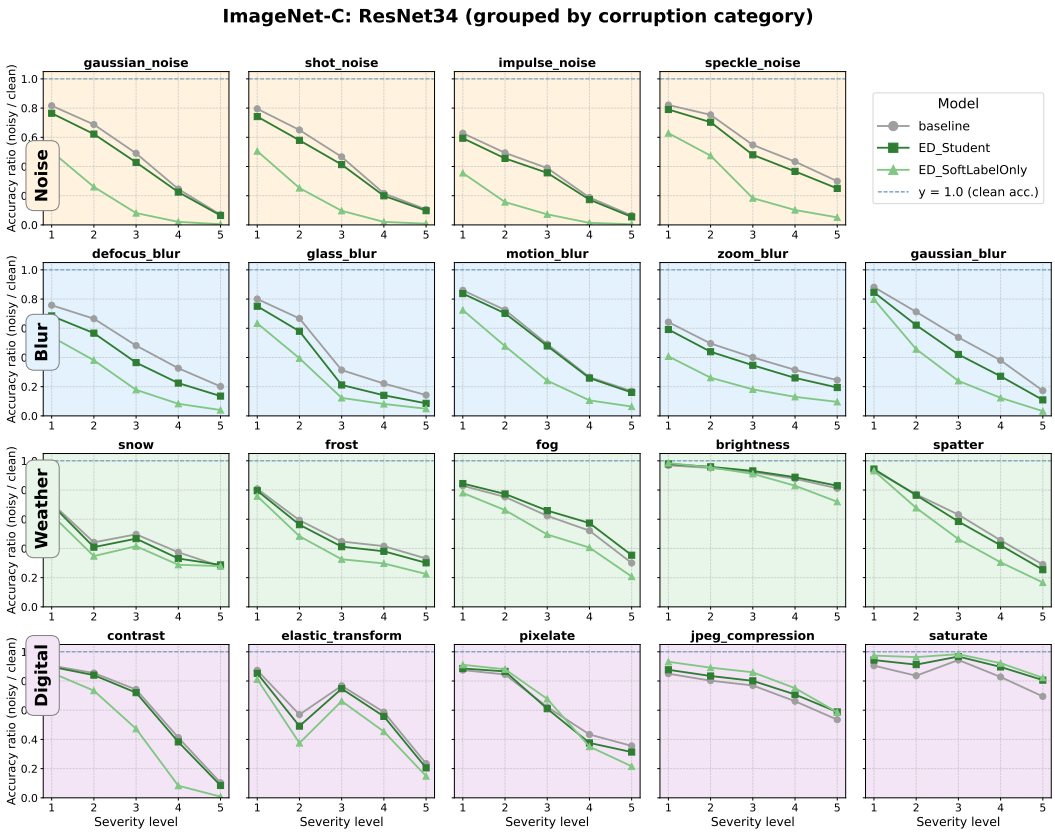


Figure 16: Top-1 accuracy of ResNet-34 models (baseline, ED_Student, ED_SoftLabelOnly) against all 19 ImageNet-C corruptions, grouped into Blur / Digital / Noise / Weather. Horizontal axis is Severity (1–5).

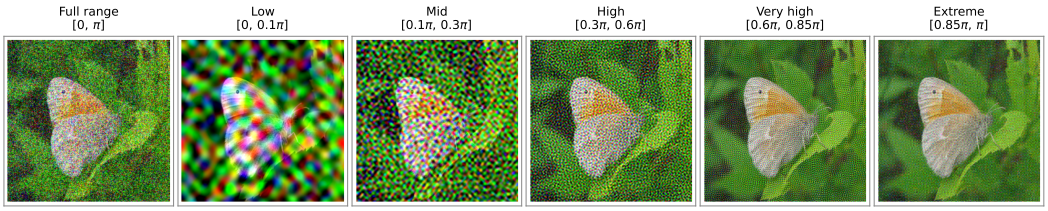


Figure 17: An ImageNet validation image with additive Gaussian Noise (Full range, $[0, \pi]$), and the results of first decomposing the noise with band-pass filters into Low ($[0, 0.1\pi]$), Mid ($[0.1\pi, 0.3\pi]$), High ($[0.3\pi, 0.6\pi]$), Very high ($[0.6\pi, 0.85\pi]$), and Extreme ($[0.85\pi, \pi]$) and then adding each band's component to the image. The Full-range noise visually appears dominated by high-frequency granular components, but it carries energy across all frequency bands.

D.4 Why Event-distilled Models are Fragile to Additive Noise

We further investigate the vulnerability to Noise corruptions confirmed in the previous subsection from the perspective of the noise's frequency characteristics. As pointed out by Yin et al. [81], additive noises such as Gaussian Noise, Shot Noise, and Impulse Noise are white noise (or close to it) in the spatial domain, and behave as broadband noise that has energy across all frequencies in the Fourier domain, from low to high.

To visually confirm this property, we first decompose Gaussian noise with band-pass filters into different frequency bands and then add each band's component to the image; the resulting images are shown in Figure 17. As the figure shows, the Full-range noise injects strong noise components into the Low ($[0, 0.1\pi]$) and Mid ($[0.1\pi, 0.3\pi]$) frequency bands on which the Event-distilled model relies for inference. As shown in the top row of Figure 18, the energy of the RGB image itself is also concentrated in the Low–Mid bands; not only do the steep edges constituting object contours appear predominantly in the mid-frequency band, but broad edge components accompanied by gradual

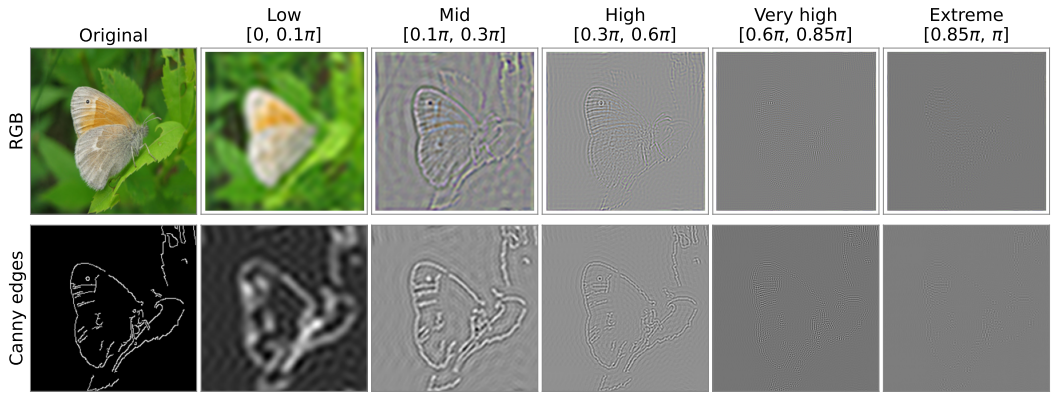


Figure 18: The same ImageNet image (RGB, top row) and its Canny edge map (Canny edges, bottom row) decomposed by band-pass filters into the same bands as in Figure 17. The Original column shows the input that each row decomposes (top: the RGB image; bottom: the Canny-processed image). The energy of the RGB image (top) is concentrated in the Low–Mid bands ($[0, 0.3\pi]$). Even the Canny map (bottom), which consists solely of edge components, similarly concentrates its energy in the Low–Mid bands and carries virtually no energy in the Very-high or higher bands. In other words, edge information is not localized in the very-high band: the low-frequency band still carries broad edge components accompanied by gradual gradients, and the energy is biased toward the mid-frequency band.

gradients are also distributed in the low-frequency band. Consequently, the bands the Event-distilled model uses to extract correct edges and the bands the noise contaminates overlap heavily.

To further isolate the frequency distribution of edge information itself, we apply the same band-pass decomposition to a Canny edge map extracted from the RGB image; the result is shown in the bottom row of Figure 18. Even though the input is composed solely of edges, its energy is similarly concentrated in the Low–Mid bands and is virtually absent from the Very-high and higher bands. This visually confirms that edge information does not have a one-to-one correspondence with very-high-frequency content: broad edge components are still distributed in the low-frequency band, while the energy is biased toward the mid-frequency band. This is consistent with our claim in this section that the Event-distilled model relies primarily on the mid-frequency band for edge information, and that contamination of the surrounding (Low–Mid) bands can therefore substantially impair performance.

The Event-distilled model has acquired a tendency to selectively exploit edge information while ignoring high-frequency texture (Section 5.1 of the main paper); accordingly, it is strongly affected by contamination of the frequency bands that can disrupt either the local or the global aspects of edge information. This is evidence, from a different angle, that the Event-distilled model is not fragile to noise in general but rather that it strongly depends on the frequency components that carry edge information, and is consistent with the observation in Appendix D.2 that performance differences disappear when noise intrudes into the frequency band on which the model relies.

E Spectral Statistics of Event vs. Natural Images

In this appendix, we corroborate the spectral trade-off hypothesis discussed in Section 5 of the main paper at the most upstream level, the frequency statistics of the data itself, by comparing the radial Power Spectral Density (PSD) of ImageNet (natural images) and N-ImageNet (event representation).

Setup. It is widely known that the PSD of natural images follows a power law $P(f) \propto 1/f^{\alpha}$ with $\alpha \approx 2$ [99]. We follow this convention and treat α as a positive exponent throughout this appendix. Following the standard procedure of van der Schaaf and van Hateren [99], Langer [100], we estimate α by linear regression in the log-log domain. We discard low-frequency components ($f < 0.02F$) that are dominated by global structure and high-frequency components ($f > 0.25F$) that are prone to noise and discretization effects, and fit only over the intermediate band, where F denotes the maximum radial frequency.

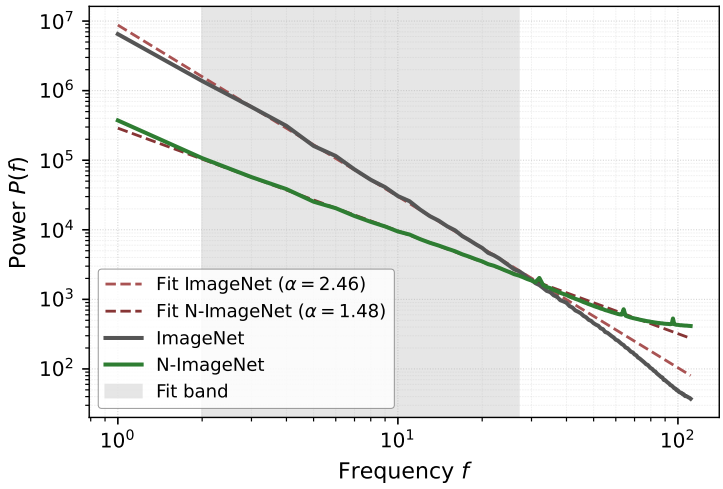


Figure 19: Radial PSD of ImageNet (natural images) and N-ImageNet (event representation) plotted in log-log space. Linear fits are computed over the intermediate frequency band ($0.02F \leq f \leq 0.25F$, gray shading; F denotes the maximum frequency) to estimate the power-law exponent α in $P(f) \propto 1/f^{\alpha}$. ImageNet yields $\alpha = 2.46$ ($R^2 = 0.9993$), broadly consistent with the empirically established $1/f^2$ statistics of natural images [99], while N-ImageNet yields $\alpha = 1.48$ ($R^2 = 0.9996$), a markedly shallower slope, indicating less concentration of energy at low frequencies.

Results and interpretation. As shown in Figure 19, ImageNet falls within the typical range of natural-image statistics ($\alpha = 2.46$), whereas N-ImageNet shows a slope shallower by roughly one point ($\alpha = 1.48$). This indicates that, in the N-ImageNet representation, the dominance of low-frequency components is weakened and the relative proportion of mid- to high-frequency components is correspondingly larger (i.e., the PSD is flatter).

This difference is consistent with the hardware design of the event sensor: an event sensor records only local changes in brightness, so smooth, spatially extended (low-frequency) textures and uniform backgrounds are not represented in its output. As a consequence, the frequency components that correspond to edges in the natural image (predominantly the mid- to high-frequency bands) are relatively emphasized, and the PSD slope becomes shallower. Note, however, that this does not mean edge information is localized exclusively in the high-frequency band: broad, diffuse edges accompanied by gradual gradients still carry energy in the low-frequency band as well (see Figure 18 in Appendix D.4). From the perspective of the spectral trade-off hypothesis discussed in Section 5 of the main paper, the N-ImageNet teacher learns from a distribution that is already richer in edge content; therefore, the student model that is distilled from it is steered, already at the input-statistics level, toward representations that do not depend on unstructured high-frequency texture and that selectively extract edge-based shape features (Section 5.1 of the main paper). The first-layer filters in Figure 3 and the spectral trade-off on ImageNet-C in Section 5.2 can be understood as the consistent downstream manifestations of this upstream input-statistics difference at the representation and behavioral levels.

F Extended Discussion of Limitations

This appendix expands on the limitations summarized in Section 7 of the main paper.

While this work provides insights regarding the inductive bias derived from event data, several limitations exist, and they suggest future research directions. First, there is a limitation arising from the data-generation process of N-ImageNet used to train the teacher: the events are obtained by recapturing RGB images displayed on a monitor. This procedure successfully teaches the model the core characteristics of the event sensor (asynchronous response to spatial brightness gradients, i.e., edges, and high temporal resolution). On the other hand, due to the constraints of the monitor, the

sensor's intrinsic high dynamic range (HDR) is not fully exploited; and because the captured content is planar imagery, complex motion parallax due to depth in the real world is missing. Second, since this validation focused on overcoming the texture bias intrinsic to CNNs, how Event distillation transforms the internal representation of Vision Transformers (ViT) [88], which are reportedly more shape-biased to begin with, is not yet known. Finally, although the validation in this work is limited to image classification, the macroscopic edge representation acquired here may function as a useful prior for dense-prediction tasks such as semantic segmentation; verifying these aspects with more native event datasets is an important step for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions (an analytical use of cross-domain distillation from event- to RGB-trained models, using the standard knowledge-distillation formulation rather than a new one; the observation of color invariance / a shift toward geometric cues / robustness to high-frequency noise and to adversarial perturbations with $0 < \epsilon \leq 1/255$; and the mechanism elucidation via the edge-based shape feature). All three are supported by the experimental results in Sections 4–6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations and Future Work" section is provided in the main paper (Section 7), with an extended discussion in Appendix F. It covers (i) the data-generation process of N-ImageNet (monitor-recapture, missing HDR / motion parallax), (ii) the focus on CNN texture bias rather than ViT, and (iii) the restriction to image classification.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper is empirical and does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 fully describes the cross-domain distillation framework and loss function (Eq. 1). Appendix A reports the teacher / student training recipes, optimizers, learning-rate schedules, augmentation pipelines, and per-architecture overrides. Training and evaluation code is provided as supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: ImageNet [6] and N-ImageNet [5] are publicly available datasets, both cited in the main paper. Training and evaluation code, together with reproduction commands and a license summary, is included as anonymized supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix A reports the data splits (Section A.1), teacher / student training recipes (Sections A.2–A.3), data-augmentation pipelines (Section A.4), the linear-probe protocol (Section A.5), the compute setup and reproduction cost (Section A.6), the additional non-event controls (Section A.8), and the SOTA training recipe (Section A.9).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to compute-budget constraints, every model in this paper is trained with a single random seed (seed=1; see Appendix A), and we therefore do not report multi-seed error bars. The qualitative robustness trends (color invariance, shape bias, high-frequency noise robustness) are nevertheless reproduced consistently across five architectures (ResNet-34, ResNet-50, VGG16-BN, ConvNeXt-Base, RepLKNet31B; Section 6.2), supporting the robustness of our findings beyond a single seed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A.6 states that all training was conducted on a single server with 8× NVIDIA A100 GPUs (Python 3.10, PyTorch 2.1.0, CUDA 12.2), and gives the full reproduction cost: 9 variants over 11 runs, about 11 days of wall-clock time in total, of which the two runs needed to verify the central claim take about two days.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that this work conforms to it. The study uses only publicly released image datasets and standard deep-learning models; no human subjects or personally identifiable information are involved.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is foundational research analyzing the inductive biases of vision models trained via cross-domain knowledge distillation. It is not tied to a specific deployable application and has no direct path to negative societal impact beyond that of any generic image-classification research.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artifacts are training code and standard image-classifier checkpoints (ResNet-34 etc.) on widely used public benchmarks. They do not include high-risk dual-use components such as generative models, large language models, or scraped web data, so no additional safeguards are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing datasets (ImageNet [6], N-ImageNet [5]), the public DiST checkpoint [5], and third-party software (PyTorch, PyTorch Lightning, OpenCV, RepLKNet [79], etc.) are credited in the main paper. The README of the supplementary code lists each dataset and dependency together with its license and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The training and evaluation code released as supplementary material is documented via a top-level README that describes the directory structure, dependencies, dataset preparation, and per-variant reproduction commands.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve human subjects research, and therefore no IRB approval was required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology of this work (cross-domain knowledge distillation between an event teacher and an RGB student) does not involve LLMs. LLMs were used only for writing and editing assistance, which per the NeurIPS LLM policy does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.