On the Reliability of Cue Conflict and Beyond

Pum Jun Kim[1,*], Seung-Ah Lee[1,*], Seongho Park[2], Dongyoon Han[3,†], and Jaejun Yoo[1,†]

[1] Ulsan National Institute of Science & Technology
[2] College of Medicine at Hanyang University
[3] NAVER AI Lab

Abstract. Understanding how neural networks rely on visual cues offers a human-interpretable view of their internal decision processes. The cue-conflict benchmark has been influential in probing shape-texture preference and in motivating the insight that stronger, human-like shape bias is often associated with improved in-domain performance. However, we find that the current stylization-based instantiation can yield unstable and ambiguous bias estimates. Specifically, stylization may not reliably instantiate perceptually valid and separable cues nor control their relative informativeness, ratio-based bias can obscure absolute cue sensitivity, and restricting evaluation to preselected classes can distort model predictions by ignoring the full decision space. Together, these factors can confound preference with cue validity, cue balance, and recognizability artifacts. We introduce `REFINED-BIAS`, an integrated dataset and evaluation framework for reliable and interpretable shape–texture bias diagnosis. `REFINED-BIAS` constructs balanced, human- and model- recognizable cue pairs using explicit definitions of shape and texture, and measures cue-specific sensitivity over the full label space via a ranking-based metric, enabling fairer cross-model comparisons. Across diverse training regimes and architectures, `REFINED-BIAS` enables fairer cross-model comparison, more faithful diagnosis of shape and texture biases, and clearer empirical conclusions, resolving inconsistencies that prior cue-conflict evaluations could not reliably disambiguate. Our code is available at `REFINED-BIAS`.

Keywords: Shape-Texture Bias · Cue Conflict Benchmark

1 Introduction

Do neural networks (NNs) exhibit biases in human-like perception, and can we diagnose them in cognitively meaningful terms? Addressing these questions is important for bridging machine and human vision, revealing how models internalize visual information and providing a basis for more human-like systems [2, 8, 11, 15, 17, 24]. Accordingly, a growing line of work has developed diagnostic benchmarks [3, 9, 18, 25, 32, 47] that evaluate model biases through a human-perceptual lens. Among them, the cue-conflict benchmark [18] remains

* Equal contributions. † Corresponding authors.

Fig. 1: Empirical instability of stylized image-based bias evaluation. (a) illustrates the core insight of cue-conflict that stronger shape bias, similar to humans, improves in-domain performance. (b) shows examples of unrecognizable cues in the cue-conflict dataset. (c) illustrates conflicting findings on this core insight of cue-conflict.

the de facto standard and the most widely adopted framework for bias analysis, consistently used in recent studies [8, 14, 24, 26–28, 34, 36, 45, 51, 52].

The cue-conflict benchmark probes cue preference using stylized cue-conflict images that combine the shape of one class with the texture of another. It introduced a core insight that has been highly influential: a human-like strong shape bias is associated with higher in-domain performance (Fig. 1a). While this idea remains highly valuable, we find that the current stylization-based instantiation yields empirically unstable evaluations. Many cue-conflict images do not clearly convey their shape and texture cue information, making them difficult to recognize for both humans and machines (Fig. 1b). Such ambiguity helps explain conflicting conclusions in recent studies on whether shape or texture drives model performance [1, 14, 30, 36] (Fig. 1c), and why the benchmark can fail to reflect explicitly shape-inducing training strategies.

These observations raise **fundamental reliability questions** for stylized cue-conflict evaluation:

Q1: *Do cue-conflict images instantiate perceptually valid and separable shape and texture cues?*

Q2: *Does the mixed-cue formulation in cue-conflict ensure controlled and balanced shape and texture cues such that both exert comparable influence?*

Q3: *Are both shape and texture cues reliably recognizable by both humans and models?*

If the answer to any of the above questions is *no*, measured "preference" can be confounded by cue construction artifacts rather than reflecting genuine perceptual bias. Here, we identify key limitations of the current cue-conflict instantiation (Fig. 2). On the construction side, stylization operationalizes shape and texture through model-dependent features, yielding imperfect cue separation and perceptual impurity (Fig. 2a), and provides no control over cue mixing ratios, often producing imbalanced cue informativeness (Fig. 2b). On the evaluation side, the ratio-based bias score obscures cue sensitivity, making models with vastly different absolute cue utilization appear similar (Fig. 2c), and restricting evaluation to a small subset of labels can distort model predictions, obscuring cue usage in the full decision space (Fig. 2d).

To address these issues, we introduce `REFINED-BIAS`, a reliable framework for integrated evaluation and a disentangled benchmark of interpretable alignment of shape and texture. We define shape as globally and locally coherent

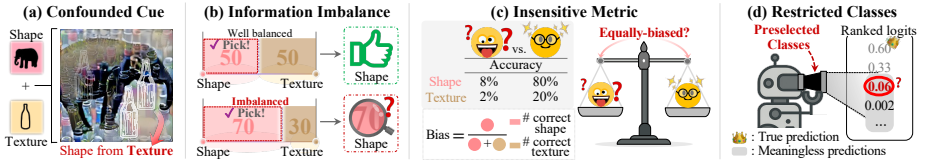

Fig. 2: Limitations of the cue-conflict benchmark: Although it has offered a valuable and well-designed framework for studying shape and texture biases, we argue that several limitations warrant further attention: (a) cue entanglement caused by stylization, where shape and texture information leak into each other, (b) imbalanced cue information caused by stylization, leading to unfair predictive contributions, (c) ignoring differences in cue sensitivity, which prevents distinguishing models with genuine biases, and (d) evaluation restricted to preselected classes.

geometric structure and texture as scale-consistent repetitive patterns [16, 25], and construct balanced, non-overlapping cue pairs from classes that are clearly recognizable to both humans and models. Building on this dataset, we propose a cue-specific sensitivity framework, evaluated on the full label set using a ranking-based metric, enabling preference to be interpreted with absolute sensitivity.

`REFINED-BIAS` restores diagnostic faithfulness in three ways. (1) It consistently reflects the effects of shape-focused learning strategies on shape-related behavior, whereas cue-conflict exhibits only partial alignment. (2) It separates cue sensitivity from relative preference, allowing fairer cross-model comparison beyond ratio-based bias. (3) It supports more reliable conclusions about how shape and texture sensitivities jointly relate to model performance. As a result, `REFINED-BIAS` enables clearer empirical insights, including architecture-dependent shape-texture trade-offs, the role of local-to-global mechanisms in improving shape understanding, and bias analysis that extends to both large-scale and quantized vision models, while resolving prior conflicting conclusions that the original cue-conflict benchmark could not reliably disambiguate.

2 Revisiting Cue Preference Benchmarking

Before introducing our benchmark, we first revisit how cue preference has been measured in prior work and why the widely adopted cue-conflict paradigm requires careful re-examination.

2.1 The Cue-conflict Paradigm

The cue-conflict benchmark [18] measures shape-texture bias using conflicting-cue images that combine the shape of one class with the texture of another through image stylization [13]. Since its introduction, it has become the de facto benchmark for studying cue preference. Importantly, however, this stylization-based design was adopted as a pragmatic solution rather than the only valid way to measure cue preference. In the original study, Geirhos *et al.* [18] initially considered more purified stimuli, such as sketches, to isolate visual cues more explicitly. Yet these introduced a substantial domain shift for standard CNNs.

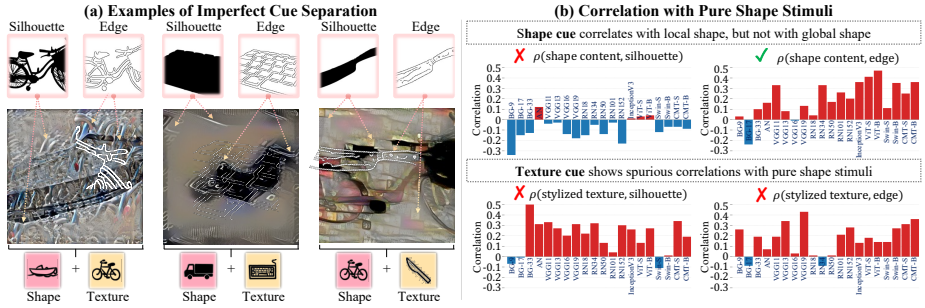

Fig. 3: Examples of imperfect cue separation in the cue-conflict dataset. (a) Qualitative examples of ambiguously extracted shape and texture cues. (b) Kendall's rank correlation of class-wise model top-1 accuracy on stylized cues and pure shape stimuli. All ImageNet-1k pretrained CNN and ViT models listed in Appendix D.1 are utilized.

Stylization was therefore adopted as a practical workaround that preserved more natural image statistics while still inducing cue conflict.

Under this implementation, shape and texture cues are operationalized via a stylization model: shape is associated with content features, and texture with style statistics such as Gram-matrix matching. Based on 16 ImageNet superclasses [19], the benchmark contains 1,280 cue-conflict images constructed from 160 shape-source images and 48 texture-source images. Not all possible combinations of shape and texture are used. Instead, a subset is selected from 7,680 (160×48) candidate pairs, and some source images are reused more frequently than others. Shape bias is then computed as $n_s/(n_t + n_s)$, where n_s and n_t denote correctly predicted shape and texture labels, respectively, and similarly for texture. While useful in practice, its current instantiation introduces several limitations that complicate reliable bias interpretation.

2.2 Limitations of the Current Cue-conflict Instantiation

Problem 1. Stylization undermines cue reliability. The cue-conflict presumes that stylization yields two perceptually valid and separable signals: shape (content) and texture (style). However, this operationalization is defined by the stylization model's internal features rather than by human perceptual criteria, and the resulting cues need not align with the intended "pure" shape and texture.

We test this assumption by comparing predictions on cue-conflict images against predictions on purified shape stimuli (silhouettes for global shape, edges for local shape). If content is a faithful shape proxy, it should correlate with both silhouette and edge performance, while style should not. Instead, as shown in Fig. 3b, content correlates mainly with edge-based signals and fails to track holistic shape, whereas stylized texture exhibits non-trivial correlations with shape proxies, indicating leakage and imperfect disentanglement. Qualitative examples further suggest that shape-like structure can remain visible in texture cues (Fig. 3a). Together, these results imply that the current stylization pipeline does not consistently instantiate clean cue separation.

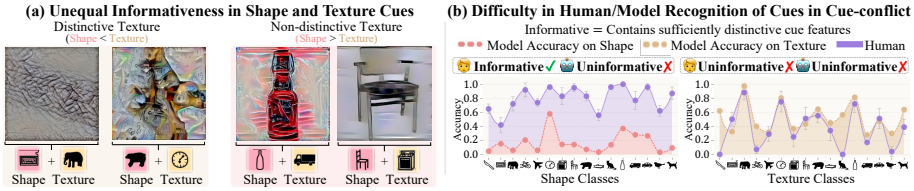

Fig. 4: Unequal recognizability of cues in the cue-conflict dataset. (a) Qualitative examples of unequal informativeness, and (b) human and model perception trends, on shape and texture cues. Model top-1 accuracies are shown with bars indicating 95% confidence intervals. All ImageNet-1k pretrained CNN and ViT models listed in Appendix D.1 are utilized. For details on the human study, see Sec. 3.1.

Even if such disentanglement were approximately valid, preference measurement requires shape and texture cues to be mixed in an equal ratio (50:50) to measure preferences fairly [3,44]. Otherwise, the resulting bias score may reflect cue imbalance rather than genuine preference. Yet, stylization offers no explicit control over the relative contribution of shape and texture, and in practice, one cue often dominates the other. As shown in Fig. 4a, a keyboard shape blended with an elephant texture contains more texture than shape, making the shape unrecognizable. Similarly, a bear shape mixed with a clock texture is not visible.

These two issues jointly lead to a third problem: the resulting cue-conflict images are often difficult even for humans to recognize in terms of both source cues. If both cues were represented faithfully, humans and models should be able to identify both the underlying shape and the transferred texture with reasonably high accuracy. Instead, as shown in Fig. 3a and Fig. 4 (see also Appendix B.1 and B.2), both human and model recognition are substantially imbalanced, with some texture classes being consistently difficult to distinguish and only a small subset remaining highly recognizable. Collectively, the current construction can confound preference with cue validity, cue balance, and recognizability artifacts.

Problem 2. Relative bias obscures cue sensitivity. Cue-conflict reports bias as a relative ratio between correct shape and texture predictions. While such a ratio can indicate directional preference, it does not capture how strongly a model actually uses either cue. For example, a model with 8% shape accuracy and 2% texture accuracy yields the same relative ratio as a model with 80% and 20%, despite the latter being far more sensitive to both cues. Thus, the relative metric conflates directional preference with absolute cue sensitivity.

This limitation is particularly problematic when the metric is used to compare models or track development progress. An increase in shape-bias score does not necessarily mean that the model has become more shape-sensitive; it may simply indicate that texture sensitivity has deteriorated more severely. Therefore, relative preference can be informative only when interpreted alongside cue-specific sensitivity, rather than as a standalone measure of cue utilization.

Problem 3. Restricted label evaluation distorts model predictions. A further limitation arises from evaluating predictions in a restricted label space. Reliable analysis should be conducted in the full decision space in which the

model was trained and performs inference. However, cue-conflict evaluation considers only a subset of candidate labels, typically the shape and texture classes of the constructed image. This filtering can distort the model's actual prediction.

For example, if a model's top-1, top-2, and top-3 predictions are "rabbit", "cat", and "dog", respectively, the true model prediction is "rabbit" (Fig. 5a). However, when the model's output space is restricted to a predefined subset of classes (*e.g.*, "cat" and "dog"), the original top-2 prediction (*i.e.*, "cat") becomes the top-1 result, thereby distorting the model's true prediction. Such distortion can mislead bias evaluation: when the model's distorted prediction happens to match the ground-truth label ("cat" in Fig. 5b), the model may appear to correctly rely on a given shape or texture cue, even though it did not genuinely recognize it. Consequently, restricted label evaluation can overestimate cue usage and obscure the model's true perceptual behavior.

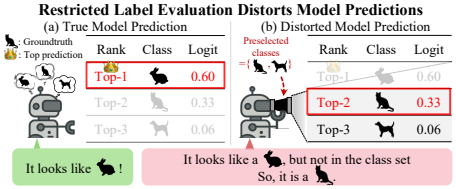

Fig. 5: Illustration of the difference between (a) true model prediction and (b) distorted model prediction. Groundtruth label is "cat", true model prediction (*i.e.*, top-1 prediction in full decision space) is "rabbit". See Appendix C.1 for more details.

2.3 Prior Critiques and the Remaining Gaps

Our critique is not the first to question aspects of the cue-conflict benchmark. Several recent studies [3, 9, 44, 47] have already pointed out important limitations in the current paradigm. Some focused on cue imbalance and attempted to construct more controlled stimuli. For example, Tartaglini *et al.* [44] fill shape silhouettes with textures, but their design overlooks local shape features and still relies on relative bias metrics, limiting fair cross-model comparison. Burgert *et al.* [3] analyze prediction shifts under cue degradation, but do not fully account for low-level texture cues and remain constrained by ratio-based bias measures. Other studies examined the limitations of relative bias metrics or evaluated cue sensitivity through degraded or texture-suppressed inputs. Wen *et al.* [47] measure sensitivity changes under global shape degradation, whereas Doshi *et al.* [9] assess shape sensitivity using texture-suppressed global shape cues.

These studies provide valuable evidence that the current cue-conflict setup should not be treated as an unquestionable standard. At the same time, prior efforts remain largely fragmented, addressing only part of the problem, such as cue balance, relative bias, or shape sensitivity, while leaving other issues unresolved. In particular, no prior work provides a unified re-examination of how cues are constructed, how predictions are evaluated, and how cue-specific sensitivity should be measured in a full-label setting. Our work is motivated by precisely this gap: rather than addressing a single limitation in isolation, we revisit cue preference benchmarking from the ground up and develop an integrated framework that improves cue construction, evaluation, and interpretability together.

Together, the issues discussed in this section suggest that cue-conflict may not fully satisfy the desiderata of a reliable bias benchmark, which leads to

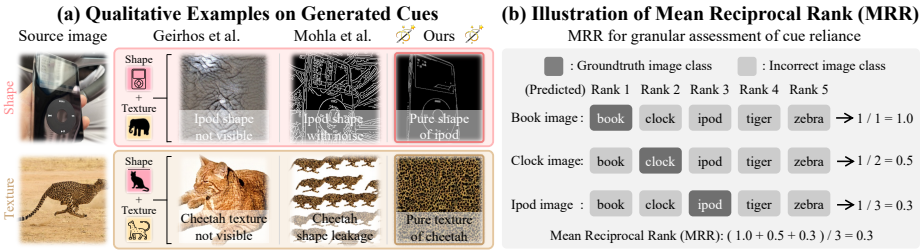

Fig. 6: `REFINED-BIAS` dataset and metric. (a) Qualitative comparison with existing cue generation methods [13,32]. (b) Computation of the mean reciprocal rank (MRR) metric. Additional qualitative examples are provided in Appendix A.2.

unintuitive evaluations in practice (Sec. 4.1 and 4.2). Existing alternatives also remain limited in delivering an intuitive model bias evaluation (Appendix C.2).

3 Methodology

We propose `REFINED-BIAS`, a new framework for accurately and reliably measuring and comparing shape and texture biases. First, we define disentangled stimuli following human perceptual standards rather than model-derived heuristics, ensuring that each cue is pure, equally informative, and clearly recognizable (Sec. 3.1). Second, we introduce a novel bias metric evaluated in the model's true decision space, capturing both preference and sensitivity (Sec. 3.2). By unifying these refined stimuli and evaluation metrics, `REFINED-BIAS` satisfies the essential desiderata for a reliable bias assessment, enabling a precise and fair comparison of perceptual capabilities across models.

3.1 Shape and Texture Cue Construction

What constitutes shape and texture in our benchmark? To address Problem 1, we define shape and texture based on human perception rather than model heuristics, and generate cues that faithfully capture these characteristics, as shown in Fig. 6a (right).

Texture is defined as a pattern that consistently repeats within patches of various image sizes. For example, classes like "honeycomb" and "dishrag" exhibit characteristic textures that remain recognizable even when divided into small patches of different sizes.

Shape, on the other hand, is defined as a non-repeating geometric structure, encompassing both global and local features [16,25]. Global geometry refers to the overall structure of an object, such as its silhouette, while local geometry includes distinctive substructures not repeated across the object. For instance, although "ipod" and "comic book" share a similar rectangular global geometry, their local structures are distinct, allowing reliable classification.

Data collection. To further mitigate the Problem 1, we construct a curated dataset of 20 ImageNet-derived superclasses (see Appendix A.2 for details), comprising 10 shape-dominant (*e.g.*, clock, hourglass) and 10 texture-dominant (*e.g.*, strawberry, brain coral) categories, selected based on human perceptual judgments. On the other side, the limited source images used in cue-conflict may introduce image-level variations, such as resolution differences that can make patterns appear more coarse or fine. To mitigate this, we collect 300 diverse images per class, 10 from ImageNet and the rest from web sources. This results in a more balanced set of shape and texture cues compared with cue-conflict, which contains only 10 shape and 3 texture sources per class. In total, our dataset contains 6,000 high-quality images, roughly five times larger than cue-conflict.

Shape and texture cues. In designing our cue dataset, we build upon the insights of Mohla *et al.* [32] rather than simply replicating their protocol, as the direct application of existing methods leaves data quality issues unresolved. As shown in Fig. 6a (middle), prior methods produce texture cues characterized by local and global shape leakage and reduced cue resolution, while shape cues suffer from noisy background clutter that obscures the object's structural integrity. Furthermore, the stylization method of Geirhos *et al.* [18] tends to inherit these inherent data issues found in the cue-conflict benchmark (Fig. 6a, left).

To avoid these issues, we design a more precise and carefully controlled cue generation pipeline. As shown in Fig. 6a (right), this pipeline ensures clearly recognizable pure shape and texture cues while preserving their full resolution. We further conduct human inspection on each generated cue image to enhance cue quality. The details on the generation pipeline are in Appendix A.1.

As shown in Fig. 7, our dataset achieves comparable shape and texture accuracies while substantially reducing class imbalance across categories. This reflects the high quality of our stimuli, which also faithfully capture structural 3D information while remaining free from the grid-like artifacts in textures (see Appendix C.3 for details). Note that the pure stimuli in Geirhos *et al.* [18] are small-scale, created for illustration rather than benchmarking, and not fully public. In contrast, our stimuli are systematically constructed and released as a public resource for reliable bias evaluation.

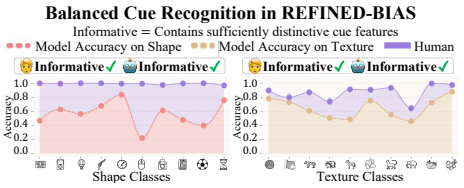

Fig. 7: Human and model perception trends on shape and texture cues of the `REFINED-BIAS` dataset. The same experimental setup as in Fig. 4 is used.

Mitigating domain shift in CNNs. To mitigate the domain shift in CNNs that often hinders bias evaluation, we carefully curate a subset of classes where either shape or texture serves as the most discriminative feature for classification. For these categories, we generate clean, artifact-free images to ensure that the intended cues remain clearly recognizable to CNNs (as well as ViTs). Consequently, `REFINED-BIAS` achieves significantly higher recognition performance across all ImageNet-1k pretrained CNNs (see Appendix D.1), with average top-1 accuracies of 46% for shape and 63% for texture. By comparison, the cue-conflict

benchmark achieves 4% (shape) and 21% (texture) in full model predictions, and 20% and 30% in distorted predictions. This indicates that `REFINED-BIAS` is significantly less susceptible to the domain shift in CNNs than cue-conflict.

Our human study details. To ensure that our shape and texture cues are clearly recognizable to humans, we conducted parallel user studies on `REFINED-BIAS` and cue-conflict. We recruited 88 participants, each completing 100 classification tasks using randomly sampled images from each benchmark. These tasks were evenly divided into shape and texture sections, where participants identified the target class of a single image for each task. To measure how consistently participants identified these cues, we calculated inter-human agreement using Fleiss' kappa (κ) [12]. As shown in Tab. 1, while `REFINED-BIAS` achieved near-perfect agreement for shape ($\kappa = 0.98$) and substantial agreement for texture ($\kappa = 0.79$), cue-conflict showed significantly lower consistency, particularly for texture ($\kappa = 0.29$), indicating inherent ambiguity in its signals. These suggest that `REFINED-BIAS` provides more recognizable cues to human evaluators.

Table 1: Inter-rater agreement scores. Fleiss' Kappa (κ) measures the degree of agreement among multiple raters in their predictions. Higher scores indicate stronger consistency in predictions and reflect how clearly cues are perceived.

Dataset	Cue Type	Fleiss' Kappa
`REFINED-BIAS`	Shape	0.9836
	Texture	0.7973
Cue-conflict	Shape	0.7276
	Texture	0.2937

3.2 Model Comparisons with Redefined Bias

Addressing the Problem 2 and Problem 3, we introduce a new metric that operates on the full logits and enables more fair cross-model comparisons. The core insights behind our metric are as follows: (1) existing bias metrics are often inflated because they rely directly on accuracy for cues, and (2) accuracy appears in both the numerator and denominator, which could distort the ratio. To resolve these issues, we (1) replace accuracy with a ranking-based metric, and (2) apply the metric only in the denominator. While various ranking-based metrics could be used, we employ Mean Reciprocal Rank (MRR) [46], which aligns well with our objectives. While accuracy assigns 0 to both 2nd and 100th place predictions, MRR distinguishes them with values of $1/2$ and $1/100$, enabling a more precise assessment of how models prioritize different cues.

Redefined Bias. Specifically, our metric computes the reciprocal ranks of the correct shape and texture labels within the model's full prediction ranking (Fig. 6b). We refer to these two components as Shape-Sens and Texture-Sens. Unlike conventional MRR, our ranking is computed over the logits:

$$\text{Shape-Sens} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_{\text{shape},i}}, \quad \text{Texture-Sens} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_{\text{texture},i}}. \tag{1}$$

Here, N is the total number of samples, $r_{\text{shape},i}$ and $r_{\text{texture},i}$ are the ranks of the correct shape and texture labels for the i-th sample in the model's ranked predictions, respectively. The relative bias for shape and texture is defined as:

$$\text{Shape preference} = \text{Shape-Sens}/(\text{Shape-Sens} + \text{Texture-Sens}), \quad (2)$$

$$\text{Texture preference} = \text{Texture-Sens}/(\text{Shape-Sens} + \text{Texture-Sens}). \quad (3)$$

In the following sections, we demonstrate that our dataset and metric effectively distinguish bias differences and identify models with genuinely strong biases.

4 Experiments

Our experimental agenda is guided by two goals. First, we validate the *correctness* of `REFINED-BIAS`, which can reliably contrast shape-texture bias across a diverse spectrum of training strategies (Sec. 4.1), aligning with our prior understanding and intuition. Building on this foundation, we further investigate how such bias varies across different model architectures (Sec. 4.2). We note that the term "bias" is used to refer to both preference and sensitivity collectively.

4.1 Validating `REFINED-BIAS` Benchmark

The credibility of all subsequent experiments hinges on the correctness of the `REFINED-BIAS` benchmark; we evaluate whether the outcomes are consistent with our intuition and whether they remain plausible given existing knowledge. To this end, we evaluate the dataset using extensive training strategies for diverse pretrained models. We then focus more on assessing the correctness of the revised sensitivity metric and examining how the resulting bias measurements correlate with ImageNet top-1 accuracy (*i.e.*, in-domain performance).

Learning strategies and hypothesis: We consider 32 ImageNet-1k pretrained models, all based on the same ResNet-50 architecture [20]. As a baseline, we use a model trained with random cropping only. Each model applies one additional training strategy on top of this baseline setup. The strategies are as follows (see Appendix D.2 and B.3 for more details and examples, respectively):

- **Shape augmentation** explicitly promotes shape-based recognition by exposing models to conflicting cues and enforcing correct shape prediction. We use three models [18, 29] following this strategy.
- **Contrastive learning** implicitly learns cue invariance by aligning representations across texture variations such as blurring, encouraging reliance on stationary shape information. We use three models [4–6] for this strategy.
- **Texture distortion** injects noise that disrupts textural information while preserving semantic structure, encouraging reliance on invariant shape features. We use five models [21–23, 31, 33] following this strategy.
- **Mixed augmentation** mixes image pairs or masks regions, enabling learning of stable shape and texture cues while reducing reliance on non-stationary cues. We use eight models [48], four with texture distortions [7].
- **Adversarial training** makes models robust to imperceptible image perturbations [14, 40]. It does not directly improve shape or texture perception, so it does not necessarily affect shape or texture bias. We consider 12 models trained with varying levels of noise.

REFINED-BIAS dataset reflects trends clearly. Based on the hypothesis, we evaluate whether each benchmark dataset faithfully reflects the expected effects of different learning strategies, using the preference metric. As shown in Tab. 2, our dataset demonstrates that shape-focused strategies consistently increase shape preference. Notably, even nuanced strategies such as mixed augmentations, which apply mild texture degradation, are accurately reflected by ours as an increased shape preference. While cue-conflict partly captures similar tendencies, many of its results are not statistically significant and show an inconsistent trend across the strategies.

Table 2: t-test on the difference between the baseline and training strategies based on the preference. The red, yellow, and gray shading indicate significant shape preference, texture preference, and no preference, respectively (α=0.05).

Model Family	Expected	Cue-conflict	REFINED-BIAS
Mixed Aug	shape	p=2.72e-04	p=0.002
Texture Dist	shape	p=0.003	p=0.010
Shape Aug	shape	p=0.181	p=0.018
Contrastive	shape	p=0.684	p=0.009
Adversarial	neither	p=5.19e-05	p=0.489

For adversarial training, results on our dataset show that robustness to imperceptible noise does not significantly affect model preference. In contrast, cue-conflict reports a larger increase in shape preference than shape-focused methods, which is counterintuitive since it is primarily aimed at improving adversarial robustness, not shape preference. Furthermore, consistent with recent findings [3] that the ImageNet-1k pretrained ResNet-50 model does not strongly rely on texture cues, our dataset indicates lower reliance on texture (0.49), while cue-conflict reports a higher texture preference (0.77). Overall, these results demonstrate that our dataset provides a more reliable reflection of model behavior.

Sensitivity metric reveals models that truly utilize cues. The primary goal of our sensitivity metric is to enable fair comparisons across models by reliably distinguishing those that utilize either shape or texture cues. To this end, we evaluate whether it can reveal cross-model differences that the preference metric misses. On our dataset, the preference metric suggests that adversarial learning induces the strongest utilization of texture cues (Fig. 8c), while on the cue-conflict dataset, it indicates the strongest utilization of shape cues (Fig. 8a). As shown in Fig. 9a and Fig. 9b, our sensitivity metric demonstrates that adversarial learning does not increase the model's utilization of either cue, whereas mixed augmentations lead models to utilize both shape and texture cues, revealing differences that the preference metric obscures. These results show that our sensitivity metric captures both cross-model differences in cue utilization and independent utilization of shape and texture cues for models that rely on both, unlike the preference metric, which fails to capture either.

Balanced cue usage positively correlates with performance. To examine the relationship between model bias and in-domain performance, we employ a dataset with pure, balanced cue information and a sensitivity metric that clearly distinguishes model differences. Following Gavrikov and Keuper [14], we fix the architecture (ResNet-50) to isolate inductive bias as a confounding factor. In Fig. 9a and Fig. 9b, our benchmark reveals that higher in-domain accuracy positively correlates with the utilization of both shape and texture cues. These

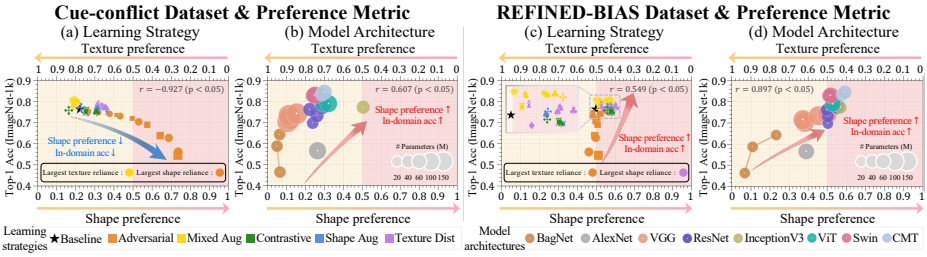

Fig. 8: Comparison of models based on preference metric and ImageNet-1k top-1 accuracy. **Varying learning strategies with a fixed architecture** are compared on the (a) cue-conflict dataset and (c) our dataset, respectively. **Varying architectures with fixed learning strategies** are compared on the (b) cue-conflict dataset and (d) our dataset, respectively. Red and yellow backgrounds indicate models with shape and texture preferences, respectively. r is the Pearson correlation coefficient.

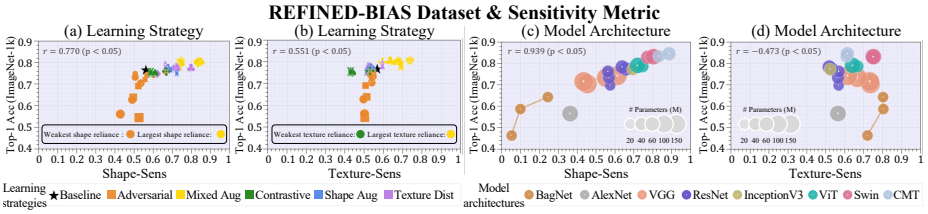

Fig. 9: Comparison of models based on the sensitivity metric and ImageNet-1k top-1 accuracy. (a) shape and (b) texture sensitivity measured across **different learning strategies with a fixed architecture**. (c) shape and (d) texture sensitivity measured across **varying architectures with fixed learning strategies**. Correlations are measured via the Pearson correlation coefficient r. See Appendix C.6 for more analysis.

results align with prior studies [29,32,43], which show that jointly utilizing shape and texture cues improves in-domain performance, further confirming that our benchmark accurately captures their complementary roles.

4.2 What Becomes Visible Once Bias Is Measured Reliably

Based on our dataset and metric, we provide an empirical analysis of how shape and texture biases vary across different model architectures and examine their relation to in-domain performance. See Appendix D.1 for full model details.

How ViT design influences cue utilization. The ViT architecture is inherently designed to capture broad, global context through its patch-wise self-attention [10], but struggles to effectively encode local features [38, 49, 50]. Subsequent designs like Swin and CMT address this by improving local-to-global feature aggregation: Swin via self-attention in progressively shifted windows, and CMT by combining local convolutional extraction with global self-attention. We investigate how this local-to-global understanding affects model behavior by analyzing shape and texture utilization using our benchmark and sensitivity metric. As shown in Fig. 9c, Swin and CMT exhibit higher shape sensitivity than ViT, indicating that improved local feature aggregation enhances shape perception

Table 3: t-test on the difference between the preference score and the no-preference baseline (*i.e.*, 0.5) across large-scale vision models. The red and gray shading indicate significant and non-significant shape preference, respectively (α=0.05).

Model	Backbone	Cue-conflict	REFINED-BIAS
OpenCLIP	ViT-B/32	p=0.276	p=0.002
	ViT-B/16	p=0.037	p=0.002
	ViT-L/16	p=0.565	p=0.001
DINOv2	ViT-S/14	p=0.030	p=3.18e-04
	ViT-B/14	p=0.011	p=0.001
	ViT-B/14	p=5.26e-21	p=3.71e-04

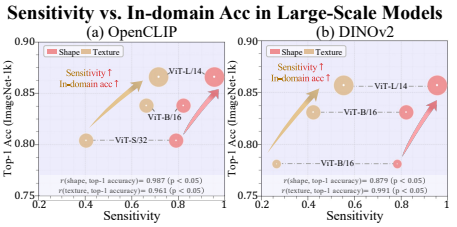

Fig. 10: Relationship between cue sensitivity and ImageNet-1k top-1 linear-probe accuracy across large-scale vision models: (a) OpenCLIP, (b) DINOv2. r denotes the Pearson correlation coefficient.

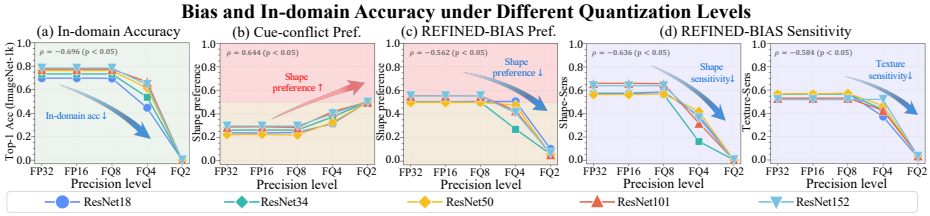

Fig. 11: Bias evaluation on quantized ImageNet-1k pretrained ResNet architectures. Relationship between quantization level and (a) ImageNet-1k top-1 accuracy, (b) preference measured on cue-conflict dataset, (c) preference, and (d) sensitivity measured on our dataset. Kendall's rank correlation ρ is computed and averaged across models.

(consistent with Shi *et al.* [42] for Swin). In contrast, Fig. 8b shows that the cue-conflict dataset, measured via the preference metric, fails to reveal this advantage: CMT shows little change, and Swin shows reduced shape preference.

Which drives performance? Shape or texture? Recent studies [1,14,30,36] using cue-conflict benchmark have yielded conflicting conclusions about whether shape preference is more beneficial for in-domain performance. Specifically, as shown in Fig. 8a, when the CNN architecture is fixed and only the training strategy varies, higher texture preference is associated with better accuracy, consistent with [14]. Conversely, when the training strategy is fixed and the architecture varies, the opposite trend emerges: models with stronger shape preference perform better (Fig. 8b), in line with [1,30,36]. These contradictory findings within the cue-conflict framework undermine a coherent understanding of preference and call into question whether the benchmark faithfully reflects its intended premise. In contrast, our dataset and preference metric provide consistent results across all setups (Fig. 8c and Fig. 8d), reliably identifying shape preference as a more important performance contributor. This consistency demonstrates that our benchmark not only accurately captures the intended insight of cue-conflict but also enables a more reliable evaluation of model preference.

What do biases reveal in large-scale models? Large-scale pretrained models such as CLIP that leverage image-text alignment are known to favor coarse semantics over pixel-level texture, and thereby encourage stronger shape re-

liance [3, 26]. This provides a natural testbed for examining two questions: Can a benchmark faithfully capture cue preferences for large models induced by different large-scale pretraining strategies, and do these preferences explain downstream recognition performance? To confirm the questions, we evaluate ViT variants from OpenCLIP [39] (pretrained on LAION-2B [41]) and DINOv2 [37] (pretrained on LVD-142M), using linear probing on ImageNet-1k with a frozen backbone and a linear classifier. In this setting, cue-conflict fails to reveal a clear shape preference across OpenCLIP models, exposing an evaluation failure mode (Tab. 3). In contrast, our benchmark consistently reveals a shape preference, providing a reliable measure of cue preference. Beyond preference, our benchmark further shows that both shape and texture sensitivities jointly drive in-domain accuracy on OpenCLIP (Fig. 10a) and DINOv2 (Fig. 10b), consistent with our findings in Fig. 9. Together, these results demonstrate that our benchmark reliably captures cue bias across diverse large-scale vision models.

What changes in bias under model quantization? Another practical case that provides a useful testbed is model quantization. Quantization progressively degrades intermediate feature representations, potentially compromising a model's visual cue perception even when semantic predictions remain relatively robust [35]. This raises the question of whether existing cue-bias benchmarks faithfully capture how visual cue processing changes under quantization. To investigate this, we analyze quantized ResNets under precisions from 32-bit to 2-bit. Under stronger quantization, cue-conflict reports an increasing shape preference (Fig. 11b), suggesting a shift toward shape reliance. However, because quantization accumulates errors throughout intermediate representations, this trend is difficult to interpret. Moreover, the reported increase in shape preference is accompanied by a decrease in in-domain accuracy (Fig. 11a), contradicting cue-conflict's central observation that stronger shape bias is associated with better recognition performance. In contrast, our benchmark shows that both shape and texture sensitivities consistently decrease as quantization becomes stronger (Fig. 11d), indicating a gradual degradation of visual cue perception, while shape preference also decreases (Fig. 11c), consistent with the drop in in-domain accuracy. These results suggest that our benchmark faithfully tracks changes in model behavior under quantization, whereas cue-conflict does not.

5 Discussion

Does `REFINED-BIAS` still evaluate cue preference? Yes, but with greater precision. By identifying which information a model can actually leverage when provided in its pure form, we derive a more grounded preference that is not confounded by the failure to recognize competing signals. This approach restores evaluative reliability while preserving the core insight of cue-conflict, enabling a more reliable assessment of how training strategies influence preference.

Could `REFINED-BIAS` still be affected by domain shift? A natural concern is that `REFINED-BIAS` may face the very domain-shift issue encountered

in the cue-conflict setting of Geirhos *et al.* [18]. Indeed, stylization-based cue-conflict benchmarks emerged in part because earlier attempts could not avoid substantial distribution mismatch when directly isolating cues. Through cleaner stimulus construction and carefully selected distinctive classes, `REFINED-BIAS` yields substantially higher top-1 accuracies, as reported in Sec. 3.1. These results suggest that `REFINED-BIAS` is less susceptible to severe domain shift than prior stylization-based cue-conflict benchmarks.

What exactly is improved by the dataset, the metric, and their combination? Our dataset resolves issues of cue purity, recognizability, and balance. It eliminates the inherent ambiguity of stylization and provides a five-times larger, scalable pool of samples that are easier for both humans and models to interpret. Our sensitivity metric enables full-label evaluation to identify genuine cue utilization. By separating "how much a model knows" from "which cue it prefers", it uniquely reveals models that genuinely utilize both cues. The integrated benefit arises from their combination, providing the empirical lesson that improved performance stems from dual reliance on both cues and that models with local-to-global attention consistently drive stronger shape utilization. Furthermore, our benchmark enables reliable bias capture in both large-scale and quantized vision models, consistently capturing model bias and its relationship to recognition performance, while faithfully reflecting changes in model behavior.

Potential limitations. While our shape cues offer diagnostic clarity, they may not fully capture 3D geometry or viewpoint dependencies. Completely isolating texture from residual shape impressions remains an open challenge, and expanding cue classes will further broaden the scope of bias analysis.

6 Conclusion

The cue-conflict benchmark has advanced our understanding of how networks use shape and texture, but its uncontrollable stylization, relative bias that obscures cue sensitivity, and strict class restrictions limit reliable bias analysis. To address these issues, we introduce `REFINED-BIAS`, a comprehensive framework designed for more controlled and precise bias evaluation. Our refined dataset and metric together provide a unified solution that addresses the limitations of cue-conflict and unresolved issues in recent benchmarks. We establish a more principled and dependable framework for evaluating cue-related biases in modern vision models.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR071940, RS-2025-02216916), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2020-II201336, the Artificial Intelligence Graduate School Program (UNIST), RS-2025-25442149), the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01).

References

1. Aygun, M., Dhar, P., Yan, Z., Mac Aodha, O., Ranjan, R.: Enhancing 2d representation learning with a 3d prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7750–7760 (2024)
2. Baker, N., Lu, H., Erlikhman, G., Kellman, P.J.: Deep convolutional networks do not classify based on global object shape. PLoS computational biology **14**(12), e1006613 (2018)
3. Burgert, T., Stoll, O., Rota, P., Demir, B.: Imagenet-trained cnns are not biased towards texture: Revisiting feature reliance through controlled suppression. arXiv preprint arXiv:2509.20234 (2025)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
5. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in Neural Information Processing Systems **33**, 22243–22255 (2020)
6. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
8. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11963–11975 (2022)
9. Doshi, F.R., Fel, T., Konkle, T., Alvarez, G.: Visual anagrams reveal hidden differences in holistic shape processing across vision models. arXiv preprint arXiv:2507.00493 (2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Finder, S.E., Amoyal, R., Treister, E., Freifeld, O.: Wavelet convolutions for large receptive fields. In: European Conference on Computer Vision. pp. 363–380. Springer (2024)
12. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
14. Gavrikov, P., Keuper, J.: Can biases in imagenet models explain generalization? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22184–22194 (2024)
15. Gavrikov, P., Keuper, J., Keuper, M.: An extended study of human-like behavior under adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2361–2368 (2023)
16. Ge, Y., Xiao, Y., Xu, Z., Wang, X., Itti, L.: Contributions of shape, texture, and color in visual recognition. In: European Conference on Computer Vision. pp. 369–386. Springer (2022)

17. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. Advances in Neural Information Processing Systems **34**, 23885–23899 (2021)
18. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2018)
19. Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. Advances in Neural Information Processing Systems **31** (2018)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
22. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
23. Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: Pixmix: Dreamlike pictures comprehensively improve safety measures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16783–16792 (2022)
24. Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. Advances in Neural Information Processing Systems **33**, 19000–19015 (2020)
25. Hermann, K., Lampinen, A.: What shapes feature representations? exploring datasets, architectures, and training. Advances in Neural Information Processing Systems **33**, 9995–10006 (2020)
26. Hernández-Cámara, P., Jaén-Lorites, J.M., Gómez-Villa, A., Vila-Tomás, J., Laparra, V., Malo, J.: On the dynamic evolution of clip texture-shape bias and its relationship to human alignment and model robustness. arXiv preprint arXiv:2508.09814 (2025)
27. Jeon, S., Kim, Y., Hong, S., Kim, J.S.: L1-induced activation sparsity shifts cnn bias from texture to shape. In: 2025 IEEE/IEIE International Conference on Consumer Electronics-Asia (ICCE-Asia). pp. 1–4. IEEE (2025)
28. Li, T., Wen, Z., Li, Y., Lee, T.S.: Emergence of shape bias in convolutional neural networks through activation sparsity. Advances in Neural Information Processing Systems **36**, 71755–71766 (2023)
29. Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., Xie, C.: Shape-texture debiased neural network training. arXiv preprint arXiv:2010.05981 (2020)
30. Lonnqvist, B., Scialom, E., Gokce, A., Merchant, Z., Herzog, M.H., Schrimpf, M.: Contour integration underlies human-like vision. arXiv preprint arXiv:2504.05253 (2025)
31. Modas, A., Rade, R., Ortiz-Jiménez, G., Moosavi-Dezfooli, S.M., Frossard, P.: Prime: A few primitives can boost robustness to common corruptions. In: European Conference on Computer Vision. pp. 623–640. Springer (2022)
32. Mohla, S., Nasery, A., Banerjee, B.: Teaching cnns to mimic human visual cognitive process & regularise texture-shape bias. In: ICASSP 2022-2022 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1805–1809. IEEE (2022)
33. Müller, P., Braun, A., Keuper, M.: Classification robustness to common optical aberrations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3632–3643 (2023)
34. Mummadi, C.K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V., Metzen, J.H.: Does enhanced shape bias improve neural network robustness to common corruptions? arXiv preprint arXiv:2104.09789 (2021)
35. Nagel, M., Amjad, R.A., Van Baalen, M., Louizos, C., Blankevoort, T.: Up or down? adaptive rounding for post-training quantization. In: International conference on machine learning. pp. 7197–7206. PMLR (2020)
36. Oliveira, T., Marques, T., Oliveira, A.L.: Connecting metrics for shape-texture knowledge in computer vision. arXiv preprint arXiv:2301.10608 (2023)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 367–376 (2021)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
40. Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust imagenet models transfer better? Advances in Neural Information Processing Systems **33**, 3533–3545 (2020)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
42. Shi, R., Li, T., Zhang, L., Yamaguchi, Y.: Visualization comparison of vision transformers and convolutional neural networks. IEEE Transactions on Multimedia **26**, 2327–2339 (2023)
43. Tang, J., Wang, G., Zhang, R., Ji, X.: Enhancing shape bias for object detection. Neurocomputing p. 132931 (2026)
44. Tartaglini, A.R., Vong, W.K., Lake, B.M.: A developmentally-inspired examination of shape versus texture bias in machines. arXiv preprint arXiv:2202.08340 (2022)
45. Tu, W., Deng, W., Gedeon, T.: Toward a holistic evaluation of robustness in clip models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
46. Voorhees, E.M., et al.: The trec-8 question answering track report. In: Trec. vol. 99, pp. 77–82 (1999)
47. Wen, Z., Li, T., Jing, Z., Lee, T.S.: Does resistance to style-transfer equal global shape bias? measuring network sensitivity to global shape configuration. arXiv preprint arXiv:2310.07555 (2023)
48. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
49. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22–31 (2021)

50. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)
51. Zhang, T., Zhu, Z.: Interpreting adversarially trained convolutional neural networks. In: International conference on machine learning. pp. 7502–7511. PMLR (2019)
52. Zhao, J., Li, T., Jiang, D., Wu, S., Ramirez, A., Lee, T.S.: Perceptual inductive bias is what you need before contrastive learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9621–9630 (2025)