

## A Notation of Feature Suppression Metrics

To quantify the effect of feature suppression, we define a transformation  $T_\tau : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{H \times W}$  that maps an original grayscale image  $\mathbf{x}$  to its feature-suppressed version  $\hat{\mathbf{x}} = T_\tau(\mathbf{x})$ , where  $\tau \in \{\text{texture}, \text{shape}\}$ . We then compare  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  using a set of similarity metrics  $\phi(\cdot)$  that target either texture or shape characteristics.

**(1) Local Variance (LV).** This metric quantifies local contrast variability and serves as a proxy for fine-grained texture information. We compute the mean variance over non-overlapping windows  $\mathbf{w}_{i,j}$  of size  $k \times k$ :

$$\phi_{\text{var}}(\mathbf{x}) = \frac{1}{N} \sum_{i,j} \text{Var}(\mathbf{w}_{i,j}), \quad (1)$$

$$\text{LV}(\mathbf{x}, \hat{\mathbf{x}}) = \min \left( 1, \frac{\phi_{\text{var}}(\hat{\mathbf{x}})}{\phi_{\text{var}}(\mathbf{x})} \right). \quad (2)$$

**(2) High-Frequency Energy Ratio (HFE).** This metric captures the spectral energy in high frequencies and is used to measure texture preservation. Using the 2D Fourier transform  $\mathcal{F}(\cdot)$ , we compute:

$$\phi_{\text{hf}}(\mathbf{x}) = \frac{\sum_{(u,v) \in \mathcal{H}} |\mathcal{F}(\mathbf{x})_{u,v}|^2}{\sum_{(u,v)} |\mathcal{F}(\mathbf{x})_{u,v}|^2}, \quad (3)$$

$$\text{HFE}(\mathbf{x}, \hat{\mathbf{x}}) = \min \left( 1, \frac{\phi_{\text{hf}}(\hat{\mathbf{x}})}{\phi_{\text{hf}}(\mathbf{x})} \right), \quad (4)$$

where  $\mathcal{H}$  is the set of frequency components with distance greater than radius  $r$  from the center.

**(3) Edge Structural Similarity (ESSIM).** This metric evaluates the similarity of edge structures, capturing shape information. Sobel gradients are computed with kernel size  $k$ :

$$\phi_{\text{sobel}}(\mathbf{x}) = \sqrt{(\partial_x \mathbf{x})^2 + (\partial_y \mathbf{x})^2}, \quad (5)$$

$$\text{ESSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \text{SSIM}(\phi_{\text{sobel}}(\mathbf{x}), \phi_{\text{sobel}}(\hat{\mathbf{x}})). \quad (6)$$

**(4) Gradient Correlation (GC).** This metric compares first-order gradients along both axes and targets shape preservation. It is defined as:

$$g_x(\mathbf{x}) = \frac{\partial \mathbf{x}}{\partial x}, \quad g_y(\mathbf{x}) = \frac{\partial \mathbf{x}}{\partial y}, \quad (7)$$

$$\text{GC}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{2} [\text{corr}(g_x(\mathbf{x}), g_x(\hat{\mathbf{x}})) + \text{corr}(g_y(\mathbf{x}), g_y(\hat{\mathbf{x}}))]. \quad (8)$$

**Averaged Texture and Shape Metrics.** We aggregate individual metrics into a Texture score and a Shape score using the *arithmetic mean*:

$$\text{Texture} = \frac{1}{2} (\text{LVS} + \text{HF}), \quad (9)$$

$$\text{Shape} = \frac{1}{2} (\text{EdgeSSIM} + \text{GC}). \quad (10)$$

All metrics are bounded in  $[0, 1]$ , where higher values indicate greater similarity to the original image and hence lower suppression of the respective feature. We set all hyperparameters to  $w = r = k = 11$ .

## B Notation for Relative Accuracy

To ensure comparability across datasets with different numbers of classes and baseline accuracies, we standardize performance using a linear rescaling:

$$\text{RelativeAccuracy} = \frac{A_{\text{sup}} - A_{\text{chance}}}{A_{\text{orig}} - A_{\text{chance}}}, \quad (11)$$

where  $A_{\text{sup}}$  denotes the accuracy under feature suppression,  $A_{\text{orig}}$  the accuracy on the original (unsuppressed) images, and  $A_{\text{chance}} = \frac{1}{C}$  the chance-level accuracy for  $C$  classes in the dataset. This mapping assigns 0 to chance-level accuracy and 1 to original accuracy, allowing direct comparison of suppression sensitivity across datasets and domains. For multi-label classification tasks, we estimate  $A_{\text{chance}}$  by simulating random predictions and computing the expected mean average precision.

## C Visual Examples for Suppression Transformations

Visual examples of the applied feature suppression transformations are provided in Figure 6 and Figure 7 illustrating images from the entry-level categories dog and bird in the ImageNet validation set. Additional examples for the RS domain are shown in Figure 16 and Figure 17, corresponding to the classes farmland and beach from the AID dataset. For the MI domain, Figure 18 and Figure 19 present examples from the classes melanocytic nevi in DermaMNIST and neutrophil in BloodMNIST.

## D Ablation Study for Suppression Effects of Transformations

To validate the robustness of our metric-based evaluation, we ablate the kernel size for all texture suppression transformations across the same 800 ImageNet images from the validation set. In this extended analysis, we include Bilateral filtering, Gaussian blur, Non-Local Means Denoising (denoted as NLMeans), Box blur, and Median filtering. For transformations with two parameters (e.g.,  $\sigma$  and  $k$  for Gaussian blur), we vary both parameters along a diagonal correspondence e.g.,  $(\sigma, k) = (0.66, 5), (1.0, 7), \dots, (2.33, 15)$  for Gaussian blur, and  $(\sigma_e, k) = (50, 5), (80, 7), \dots, (200, 15)$  for Bilateral filtering. For Non-Local Means Denoising, we use a slightly offset mapping:  $(h, k) = (5, 5), (5, 7), (10, 9), \dots, (25, 15)$  where  $k$  is used as template window size.

Figure 8 presents the effects of these transformations on normalized high-frequency energy (HFE) and gradient correlation (GC). Bilateral filtering achieves the most favorable trade-off, consistently reducing HFE while preserving edge structure to a high degree across kernel sizes. Gaussian blur suppresses texture even more strongly than Bilateral filtering, but still maintains a reasonable level of shape preservation, making it a competitive alternative when stronger smoothing is required. In contrast, Median and Box blur aggressively reduce texture but at the expense of substantial degradation in edge information. Non-Local Means Denoising preserves structural information well but is comparatively less effective at suppressing texture.

## E Experimental Details for Experiment Human vs. CNNs

### E.1 Participants Instruction

#### Participant Instructions

Thank you for participating in our visual perception study. This study investigates how humans recognize objects when certain visual features are suppressed. Please read the following instructions carefully before beginning.

#### General Information

You will be shown a series of images, each belonging to one of 16 everyday object categories (e.g., cat, car, airplane). In each trial, your task is to classify the object as accurately as possible.

#### Trial Procedure

Each trial will proceed as follows:



Figure 6: Visual illustration of feature suppression transformations applied to a sample image from the ImageNet validation set belonging to the entry-level category dog. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.

- (1) A small black fixation square will appear in the center of the screen for 300 milliseconds. Please focus your gaze on it.
- (2) An image will be shown for 200 milliseconds. The image may appear altered (e.g., blurred, gray, or shuffled).
- (3) Immediately after the image, a noise mask will appear briefly to reduce visual aftereffects.

### Response Task

After each image, you will see a  $4 \times 4$  grid of category labels. Click on the label that best matches the object you saw. If you are unsure or could not recognize the object, you may select the “not clear” option. Each image will be shown only once. Please respond based on your first impression.

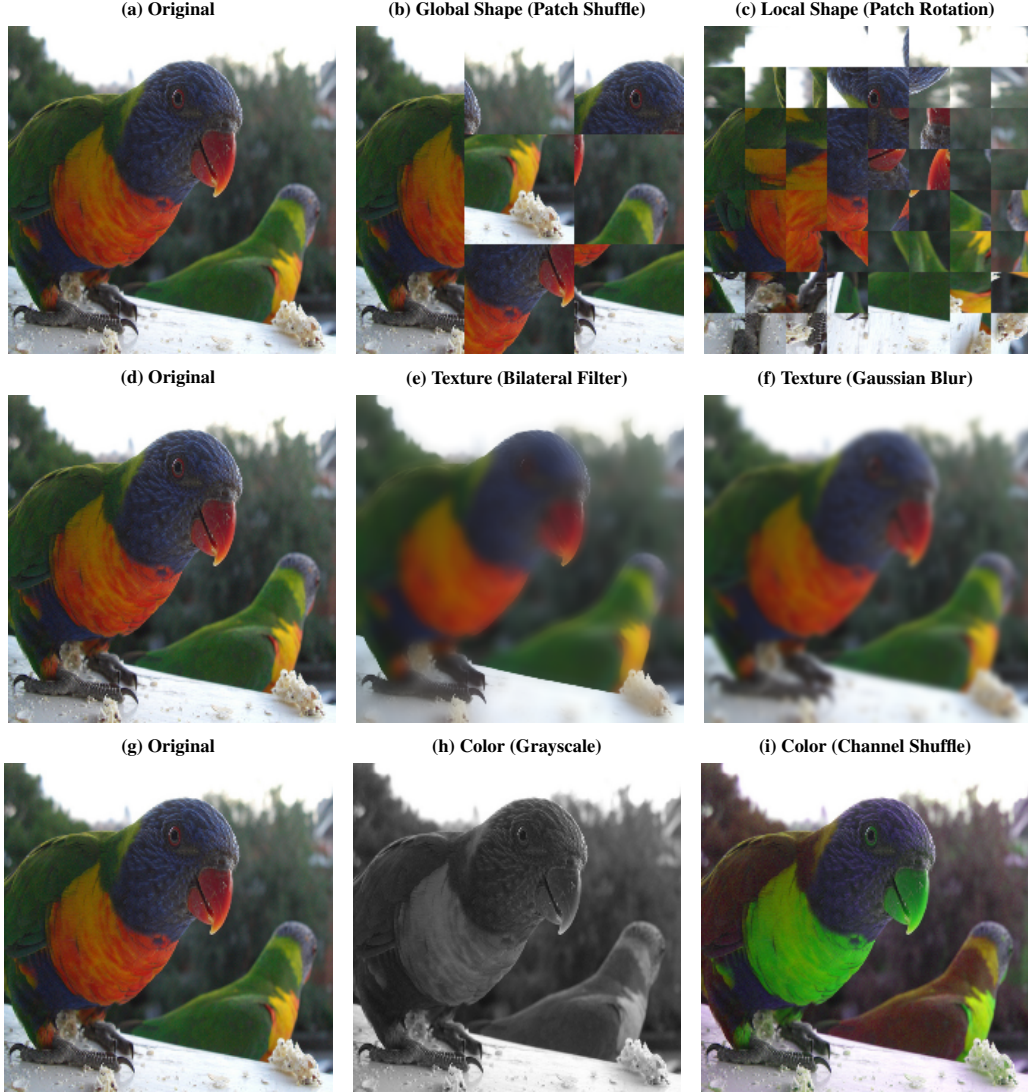


Figure 7: Visual illustration of feature suppression transformations applied to a sample image from the ImageNet validation set belonging to the entry-level category bird. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.

### Estimated Duration

The study will take approximately 35–45 minutes. Please complete it in one sitting and avoid distractions.

### Participation and Data Protection

Participation in this study is entirely voluntary. You may stop the study at any time without providing a reason and without any negative consequences. Before beginning the experiment, you will be asked to provide written informed consent. By doing so, you confirm that you understand the nature and purpose of the study and agree to participate. All data collected will be stored in an anonymized form and handled in accordance with institutional data protection policies. No personal identifying information will be published or shared.

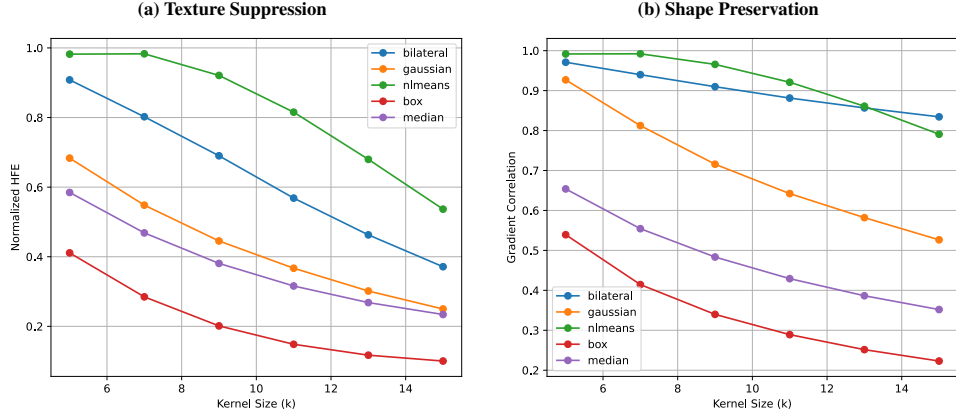


Figure 8: Ablation of kernel parameter for quantitative evaluation of feature suppression transformations. **(a)** Normalized high-frequency energy quantifies texture removal. **(b)** Gradient correlation reflects shape preservation.

By continuing, you confirm that you have read and understood the instructions.

## E.2 Attention Test

Every 100 trials, an unannounced attention test was administered (in total 7). The participant was informed via a small text box about the attention test, in which the participant was informed that they would see an object of class A in the next image and that they had to click on a different class B to successfully pass it. Only if the attention test was correctly passed, we considered the results as valid.

## E.3 Screenshots

In Figure 9, screenshots of the tool for conducting the human study are shown.

## F Statistical Significance Test for Experiment Humans vs. CNNs

Two-sided paired t-tests performed for the experiment in Section 5 reveal statistically significant differences between human and ResNet50-standard performance across all suppression conditions ( $p < 0.001$ ). Effect sizes are large in all cases (Cohen’s  $d$  ranging from 8.69 to 39.83), confirming strong and systematic divergences in feature reliance (see Table 4). The t-tests are performed using the scipy library [66].

Table 5 reports mean accuracies with 95% confidence intervals for humans and the ResNet50-standard under each suppression condition, along with the corresponding human–model accuracy differences. To assess inter-subject and inter-model consistency, we further estimate the noise ceiling: the item-wise human noise ceiling is  $0.542 \pm 0.055$ , and the model noise ceiling is  $0.646 \pm 0.015$ .

Table 4: Paired two-sided  $t$ -test comparing human vs. ResNet50 accuracy under each suppression condition ( $n = 4$  per group). Cohen’s  $d$  quantifies the standardized effect size.

Suppression Type	$t$ -statistic	$p$ -value	Cohen’s $d$	Power
Global Shape	24.76	0.0001	12.38	1.000
Local Shape	48.16	<0.0001	24.08	1.000
Texture	79.65	<0.0001	39.83	1.000
Color	17.38	0.0004	8.69	1.000



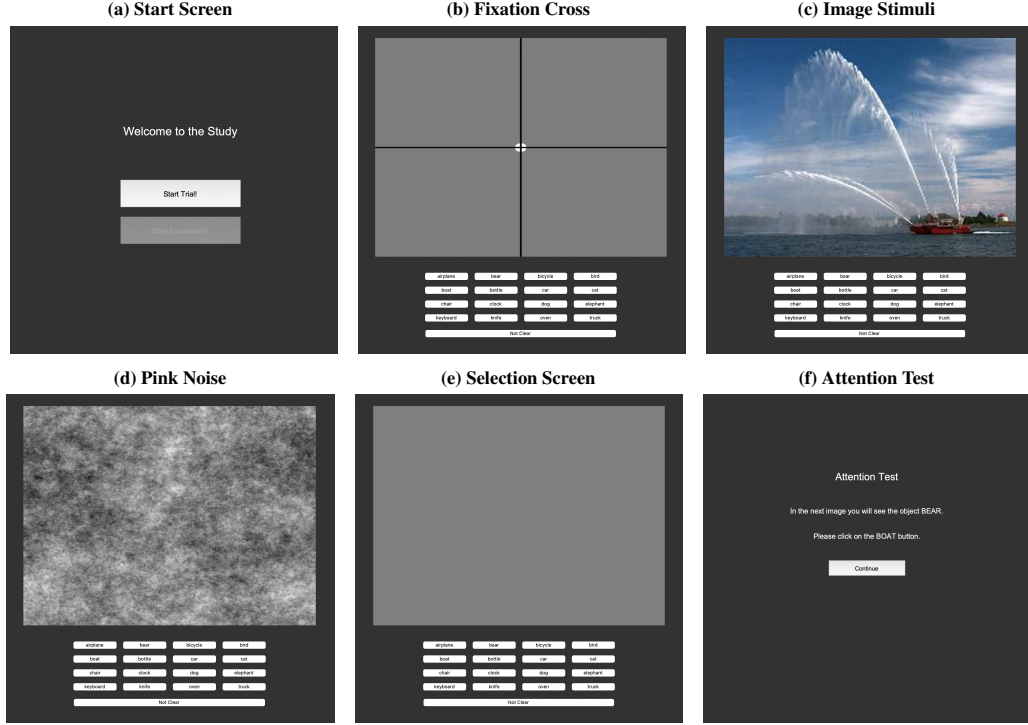


Figure 9: Screenshots from the human study.

Table 5: Human and ResNet50-standard accuracies under different feature suppression conditions. Values denote mean accuracies with 95% confidence intervals (CI), and the rightmost column reports the human–model accuracy difference.

Suppression Type	Human Accuracy (CI)	ResNet Accuracy (CI)	Difference (CI)
Global Shape	0.965 ( $\pm 0.0067$ )	0.832 ( $\pm 0.0102$ )	0.133 ( $\pm 0.0121$ )
Local Shape	0.763 ( $\pm 0.0092$ )	0.276 ( $\pm 0.0141$ )	0.487 ( $\pm 0.0168$ )
Texture	0.979 ( $\pm 0.0064$ )	0.795 ( $\pm 0.0073$ )	0.184 ( $\pm 0.0097$ )
Color	0.999 ( $\pm 0.0052$ )	0.924 ( $\pm 0.0035$ )	0.075 ( $\pm 0.0063$ )

## G Comparison of Softmax Thresholding and Argmax Decision Rules

In Table 6 we compare our heuristic decision rule used in the main experiments, which aggregates subclass probabilities via summed softmax and applies a 0.5 threshold, with a plain argmax rule. Argmax increases absolute accuracy slightly, yet the relative degradation patterns across suppression types and models remain stable, indicating that the main conclusions about feature reliance are robust to the choice of decision rule.

## H Absolute Performance of Models trained on CV, MI and RS

Table 7 summarizes the maximum validation performance of ResNet50 across all datasets and domains. Results are reported as macro and micro accuracy, separately for models trained from scratch and, where applicable, fine-tuned from ImageNet pretrained weights. For DeepGlobe, we report performance in mean average precision.

Table 6: Comparison of softmax thresholding and argmax decision rules. Values report accuracy under each suppression condition and on original images for three representative models.

Model variant	Global Shape	Local Shape	Texture	Color	Original
ResNet50-standard (Sum + Softmax > 0.5)	0.832	0.276	0.795	0.924	0.954
ResNet50-standard (Argmax)	0.880	0.361	0.840	0.944	0.962
ResNet50-sota (Sum + Softmax > 0.5)	0.943	0.618	0.867	0.948	0.931
ResNet50-sota (Argmax)	0.949	0.665	0.904	0.965	0.940
ConvNeXtV2 (Sum + Softmax > 0.5)	0.949	0.647	0.925	0.969	0.940
ConvNeXtV2 (Argmax)	0.979	0.748	0.956	0.984	0.944

Table 7: Maximum validation performance (macro and micro accuracy) of ResNet50 across domains, datasets, and training settings. For DeepGlobe, the performance metric is mean average precision.

Domain	Dataset	Training Type	Accuracy Macro	Accuracy Micro
Computer Vision	Caltech101	Pretrained	0.9523	0.9700
Computer Vision	STL10	Pretrained	0.9800	0.9800
Computer Vision	OxfordIIITPet	Pretrained	0.9404	0.9402
Computer Vision	Flowers102	Pretrained	0.9225	0.9225
Computer Vision	ImageNet	From Scratch	0.7423	0.7423
Computer Vision	Caltech101	From Scratch	0.7012	0.7972
Computer Vision	STL10	From Scratch	0.7800	0.7800
Computer Vision	OxfordIIITPet	From Scratch	0.6207	0.6214
Computer Vision	Flowers102	From Scratch	0.5186	0.5186
Medical Imaging	BloodMNIST	From Scratch	0.9868	0.9848
Medical Imaging	DermaMNIST	From Scratch	0.5444	0.7906
Medical Imaging	PathMNIST	From Scratch	0.9983	0.9984
Medical Imaging	ChestMNIST	From Scratch	0.7048	0.7048
Medical Imaging	RetinaMNIST	From Scratch	0.4446	0.5750
Remote Sensing	AID	From Scratch	0.8812	0.8847
Remote Sensing	PatternNet	From Scratch	0.9921	0.9921
Remote Sensing	RSD46WHU	From Scratch	0.8123	0.8177
Remote Sensing	UCMerced	From Scratch	0.9335	0.9335
Remote Sensing	DeepGlobe	From Scratch	0.8857	0.9295

## I Additional Experiments

### I.1 Class-wise analysis for ImageNet in Experiment II

To examine whether the feature reliance patterns observed in Experiment II generalize across categories or are driven by a small subset of classes, we conduct a class-wise analysis on ImageNet-1K. We evaluate performance under local shape suppression (Patch Shuffle, grid size 6), texture suppression (bilateral filter, kernel size 12), and color suppression (grayscale). For each of the 1000 ImageNet classes, we computed relative accuracy under suppression and visualized the distributions using kernel density estimates (KDEs). The class-level distributions in Figure 10 confirm the global trend: scores under texture suppression are clearly shifted toward higher values (mean = 0.62) compared to local shape suppression (mean = 0.24), with only modest overlap between the distributions (approximately 25–30%).

To assess consistency across classes, we computed the proportion of categories following the global ranking:

- 88.0% of classes show greater reliance on local shape than on texture

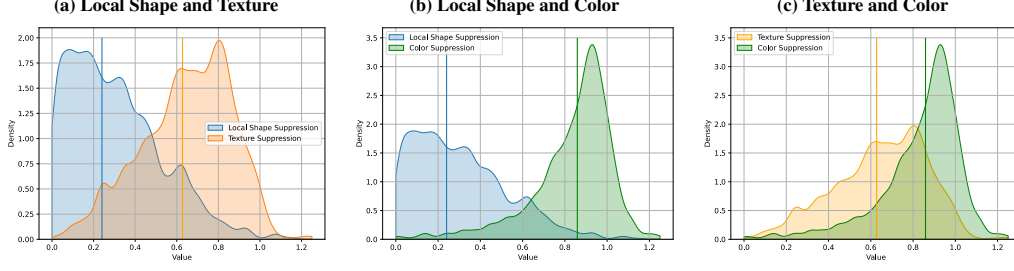


Figure 10: Kernel density estimates of class-wise relative accuracy under local shape, texture, and color suppression for ImageNet-1K.

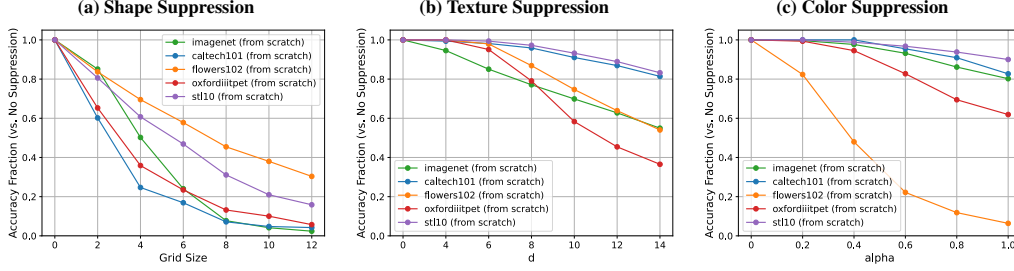


Figure 11: Feature suppression results on CV datasets for a ResNet50 trained from scratch. (a) Shape suppression via Patch Shuffle. (b) Texture suppression via bilateral filtering. (c) Color suppression via grayscale.

- 94.2% show greater reliance on local shape than on color
- 77.6% show greater reliance on texture than on color

These results demonstrate that the observed reliance patterns hold broadly across the dataset rather than being driven by isolated outlier categories. To complement the quantitative analysis, we identified representative outlier classes. Representative outlier classes with unusually low reliance on local shape include *corn*, *Appenzeller (cheese)*, *bookstall*, *spider's web*, *zebra*, *guacamole*, and *stone wall*. Classes with unusually high texture reliance include *Chesapeake Bay retriever*, *Crotalus cerastes*, *stingray*, *bath towel*, *Alligator mississippiensis*, and *bolete*. Outliers with stronger color reliance include *sunglass*, *tank suit*, *ladle*, *coffeepot*, *chain*, and *tam-tam*.

## I.2 CV Feature Reliance for Models trained from Scratch

Figure 11 shows suppression results for CV datasets using models trained from scratch. Compared to their fine-tuned counterparts, these models exhibit greater sensitivity to color and texture suppression, indicating higher reliance on these features. Interestingly, relative accuracy under light shape suppression (e.g., Patch Shuffle with small grid size) is lower than for fine-tuned models, while performance under strong shape suppression improves in several cases, indicating that these models increasingly use texture or color cues when shape information is heavily degraded. This effect is most pronounced for Flowers102, where the model retains comparatively high accuracy despite aggressive shape perturbation.

## I.3 MI and RS Feature Reliance for ImageNet-pretrained models

Figure 12 reports suppression results for MI and RS datasets using ImageNet-pretrained ResNet50 backbones fine-tuned on the respective training sets. The experimental setup and suppression protocols mirror those of Section 5.2 in the main paper. Across both domains, the overall reliance profiles remain broadly consistent with the models trained from scratch. However, pretraining introduces systematic shifts. In RS, models exhibit slightly stronger shape reliance and reduced sensitivity to texture and color suppression, in some cases up to 7% greater degradation under shape suppression and up to 20% less degradation under texture or color suppression. This aligns with the



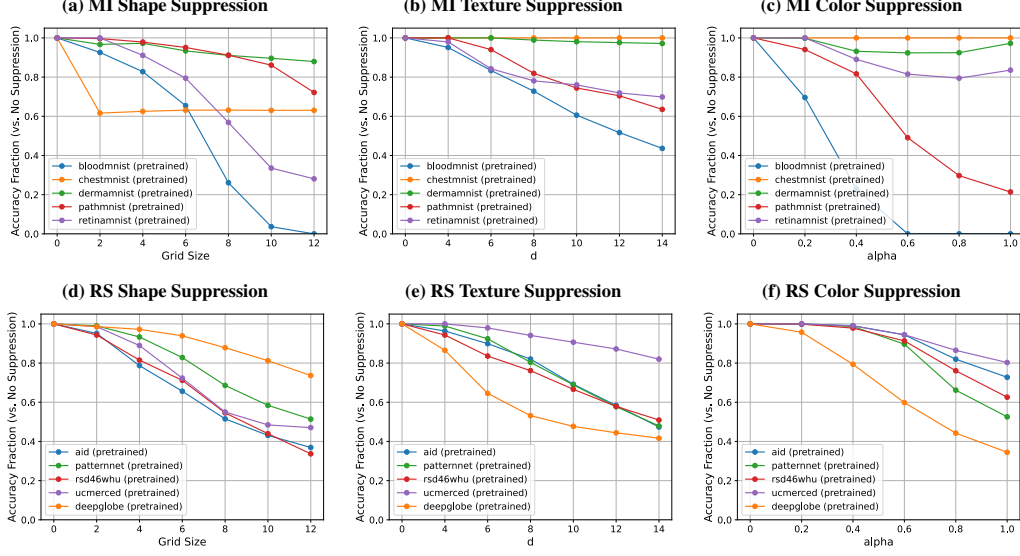


Figure 12: Feature suppression results across MI and RS domains when pre-trained on ImageNet. **Top row (a–c)**: ResNet50 pretrained on ImageNet and fine-tuned on MedMNIST-v2 (medical imaging). **Middle row (d–f)**: ResNet50 pretrained on ImageNet and fine-tuned on high-resolution RS datasets. Columns correspond to: **(a, d)** shape suppression (Patch Shuffle), **(b, e)** texture suppression (Bilateral Filter), and **(c, f)** color suppression (Grayscale).

trends observed in CV, where ImageNet pretraining enhances robustness to non-shape perturbations. In MI, the effects are more heterogeneous. For BloodMNIST, pretraining increases sensitivity to shape suppression, particularly under strong perturbations (e.g., Patch Shuffle with grid size 8), while for PathMNIST it amplifies the impact of texture suppression.

#### I.4 Alternative Transformations for Domain-specific Feature Reliance

Figure 13 presents the results using alternative transformations targeting the same feature types: Patch Rotation for shape suppression, Gaussian Blur for texture suppression, and Channel Shuffle for color suppression. Overall, the results strongly correlate with those reported in the main paper using the primary suppression transformations. The only notable deviations occur for DermaMNIST and BloodMNIST, which show increased robustness to shape suppression via Patch Rotation, and for DermaMNIST, which exhibits a stronger decline under color suppression via Channel Shuffle. Across all other datasets and domains, the relative suppression effects remain consistent.

#### I.5 Joint Suppression of Multiple Feature Types

The potential interdependence between features raises the question of whether the observed reliance patterns persist when multiple features are suppressed simultaneously. To probe this, we conduct an additional experiment on the CV datasets when pretrained on ImageNet using joint suppression. We evaluate three combinations: texture and color suppression (preserving only shape), shape and color suppression (preserving only texture), and shape and texture suppression (preserving only color). For each case, we measure relative accuracy across increasing suppression strengths, analogous to the procedure in Section 5.2.

The results reinforce and extend the single-feature findings. Relative accuracy is highest when only shape is preserved, lower when only texture remains, and almost lost when only color is available (see Figure 14). These outcomes suggest that the relative importance of features remains stable even under combined suppression.

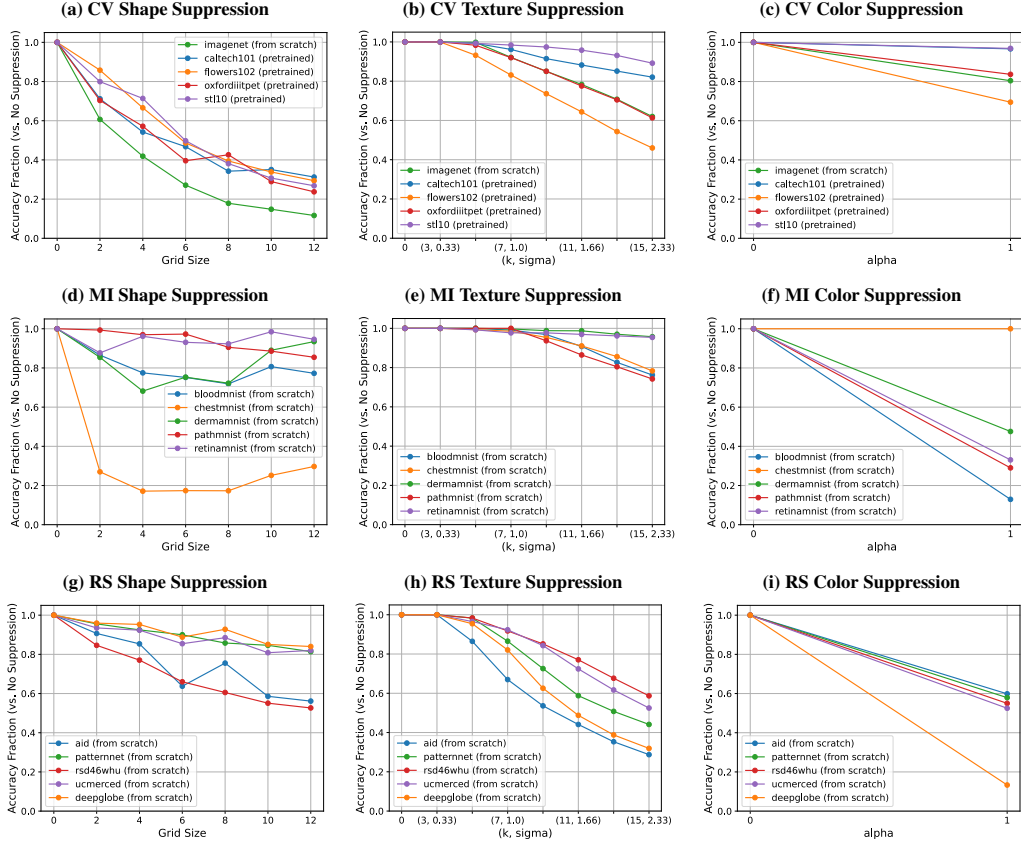


Figure 13: Feature suppression results across three domains. **Top row (a–c)**: ResNet50 pretrained on ImageNet and fine-tuned on CV datasets. **Middle row (d–f)**: ResNet50 trained from scratch on MedMNIST-v2 (medical imaging). **Bottom row (g–i)**: ResNet50 trained from scratch on high-resolution RS datasets. Columns correspond to: **(a, d, g)** shape suppression (Patch Rotation), **(b, e, h)** texture suppression (Gaussian Blur), and **(c, f, i)** color suppression (Channel Shuffle).

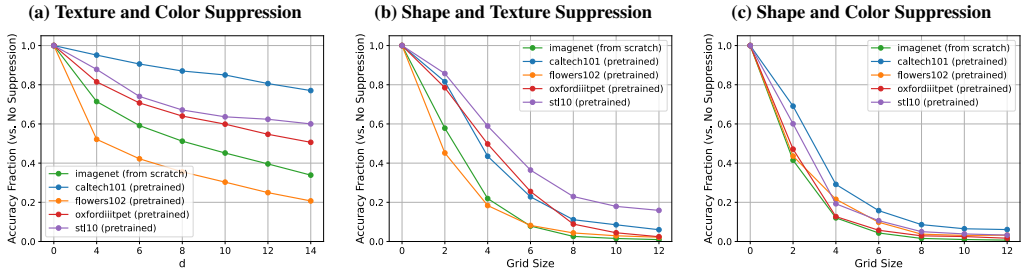


Figure 14: Joint feature suppression results for a ResNet50 pretrained on ImageNet and fine-tuned on CV datasets. **(a)** Texture suppression via bilateral filtering and color suppression via grayscale (only shape preserving). **(b)** Shape suppression via Patch Shuffle and color suppression via grayscale (only texture preserving). **(c)** Shape suppression via Patch Shuffle and texture suppression via bilateral filtering (only color preserving).

## J Control Experiment for Block-Edge Artifacts in Patch Shuffle

Patch Shuffle simultaneously disrupts local spatial continuity and introduces artificial block-edge structures. To examine whether the observed performance degradation could be attributed primarily to block artifacts, we design a control condition that isolates the grid structure from the shuffling

operation. For grid sizes of 2, 4, and 8, we generate an overlay variant of Patch Shuffle as follows. We first apply Patch Shuffle to an image, then extract the 2-pixel-wide block boundaries (1 pixel on either side). These boundaries are superimposed onto the original unshuffled image, preserving the global and local content while mimicking the block structure characteristic of Patch Shuffle. This procedure introduces visible grid lines without altering the patch arrangement. An example of the procedure is presented in Figure 15.

It is important to note that even the overlay condition introduces minor shape discontinuities, since the superimposed boundaries can slightly interrupt edge continuity. Consequently, the overlay condition still reflects two effects: (i) the presence of block edges and (ii) minor local shape disruption. Nevertheless, the stronger performance degradation observed under full Patch Shuffle compared to the overlay variant indicates (see Table 8) that block artifacts alone do not explain the results. Instead, the principal effect arises from the combined disruption of local spatial structure.

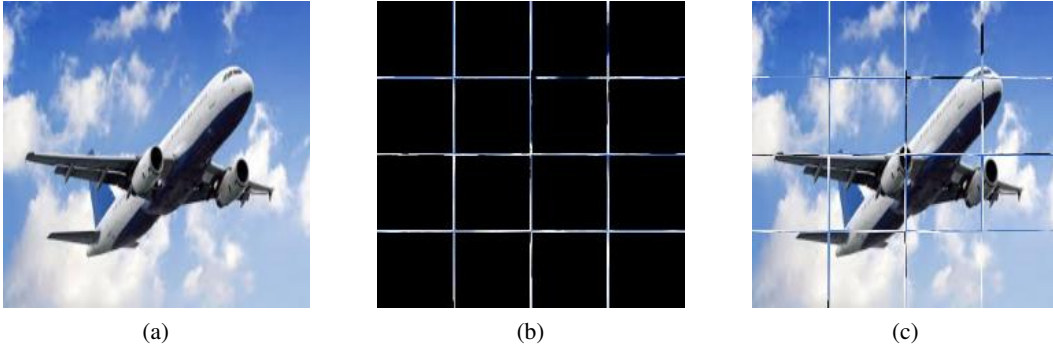


Figure 15: Example of superimposition of grid-structure for control experiment for block-edge artifacts. (a) Original image. (b) Extracted grid structure from the shuffled image. (c) Image with grid structure as overlay.

Table 8: Control experiment comparing Patch Shuffle with the grid-overlay variant, which mimics block structures without altering patch arrangement. Results are reported for different models and grid sizes.

Model	Grid Size	Overlay	Patch Shuffle
ResNet50-standard	2	0.950	0.921
	4	0.809	0.548
	8	0.540	0.069
ResNet50-sota	2	0.983	0.980
	4	0.870	0.837
	8	0.554	0.344
ConvNeXtV2	2	0.991	0.983
	4	0.957	0.859
	8	0.911	0.347

## K Implementation Details

### K.1 Timm Pretraining Hyperparameter

The following tables provide training hyperparameters for all evaluated CNNs, transformer-based models, and hybrid architectures used in Experiment I. All models, except ResNet50-standard, were obtained as pretrained checkpoints from the `timm` library and evaluated without further fine-tuning. CNNs are listed in Table 9, transformer and hybrid models in Table 10, and additional pretraining hyperparameters, if applicable, in Table 11.

We compiled the hyperparameter settings from a combination of official papers, GitHub repositories, HuggingFace model cards, and the `timm` source code. While we aim to be as faithful as possible, no

Table 9: Training hyperparameters for evaluated CNNs obtained from the timm library.

Category	ResNet50 -sota	ConvNeXt Tiny	ConvNeXt- V2-Tiny	EfficientNet -B5	EfficientNet- V2-RW-T	MobileNet V3-Large
Pretraining	–	–	IN-22k (FC-MAE)	JFT-300M (Noisy Student)	–	–
Input Resolution	224×224	224×224	224×224	456×456	224×224	224×224
Epochs	600	300	300	350	600	600
Batch Size	2048	4096	1024	2048	2048	2048
Optimizer	LAMB	AdamW	AdamW	RMSProp	RMSProp	RMSProp
Decay / $\beta_2$	–	0.999	0.999	0.999	0.9	0.9
Momentum / $\beta_1$	–	0.9	0.9	0.9	0.9	0.9
Base LR	5e-3	4e-3	8e-4	0.256	0.18	0.18
LR Schedule	Cosine	Cosine	Cosine	RMSProp decay	Step exp decay	Step exp decay
Decay Rate	–	–	–	–	0.988	0.988
Warmup Epochs	5	20	40	5	5	5
Warmup Schedule	Linear	Linear	Linear	–	–	–
Label Smoothing	0.1	0.1	0.1	0.1	0.1	0.1
RandAugment	(7, 0.5)	(9, 0.5)	(9, 0.5)	–	(8, 2, 1.0)	(8, 2, 1.0)
AutoAugment	–	–	–	yes	–	–
Mixup $\alpha$	0.2	0.8	0.8	0.2	0.2	0.2
CutMix $\alpha$	1.0	1.0	1.0	–	–	–
Rand. Erasing $p$	0.25	0.25	–	–	0.35	0.35
Dropout	–	–	–	0.2	0.2	0.2
Stoch. Depth	0.05	0.1	–	–	0.1	0.1
Drop Path	–	–	0.2	0.2	–	–
Layer-wise LR Decay	–	0.65	0.9	–	–	–
Weight Init	–	Trunc. Normal	–	Trunc. Normal	–	–
Layer Scale Init	–	1e-6	–	–	–	–
Head Init Scale	–	–	0.001	–	–	–
EMA	–	0.9999	0.9999	–	0.9999	0.9999
Loss Function	BCE	CE	CE	CE	CE	CE
Mixed Precision	Yes	Yes	Yes	Yes	Yes	Yes
Top-1 Accuracy	80.4%	82.1%	~83–84%	83.6%	79.4%	75.2%

centralized specification of all training details exists within `timm`. Accordingly, some entries (e.g., MixUp and CutMix) are marked as *yes* to indicate usage without a reported  $\alpha$  value.

## K.2 Training Hyperparameter ResNet50 from Scratch

Table [12](#) summarizes the training hyperparameters for all ResNet50 models (denoted as ResNet50-sota in Section [5](#)) trained from scratch that we used in our experiments. All models were trained or fine-tuned with minimal regularization to ensure comparability across datasets and domains. Data augmentation was limited to Random Resized Crop (RRC) and Horizontal Flip (HF). RRC was applied with a scale range of (0.3, 1.0), default aspect ratio, and probability 1.0. Through RRC all images were resized to  $224 \times 224$ . Horizontal flipping was applied with a probability of 0.5. For all computer vision datasets, ImageNet normalization statistics were used. For remote sensing and medical imaging datasets, dataset-specific statistics were computed from the training set. All models were trained using the cross-entropy loss, except for ChestMNIST (binary classification) and DeepGlobe (multi-label classification), which used binary cross-entropy. When using cosine annealing with warm restarts as learning rate scheduler, we set  $T_0 = 10$  epochs,  $\eta_{\min} = 1 \times 10^{-6}$ , and  $T_{\text{mult}} = 2$ , except for fine-tuning where  $T_{\text{mult}} = 1$ . For each dataset, the checkpoint with the highest validation accuracy was selected for subsequent suppression-based evaluation.

## K.3 Computation Resources

All experiments were conducted on an internal server equipped with 2× AMD EPYC 9554 64-core processors (256 threads), 6× NVIDIA H100 PCIe GPUs (each with 81 GB memory, CUDA 12.2), and 1.5 TiB of system RAM. The system runs Ubuntu 22.04 with Linux kernel 5.15 and NVIDIA driver version 535.183.01.

Training times for ResNet50 models varied by dataset. Training on ImageNet took approximately 10 days on a single GPU. For smaller CV datasets, training from scratch took 30 minutes for Flowers102, 120 minutes for STL-10 and Caltech101, and 90 minutes for Oxford-IIITPet. Fine-tuning on the same datasets required 10 minutes for Flowers102, 40 minutes for STL-10 and Caltech101, and 30 minutes

Table 10: Training hyperparameters for evaluated transformers and hybrid architectures obtained from the timm library. MixUp and CutMix are marked as *yes* where applied but unspecified.

Category	ConvMixer-768/32	ViT-B/16	DeiT-B	Swin-BP4-W7	CLIP ViT-B/16
Pretraining	-	ImageNet-21k	-	-	400M image-text pairs
Input Resolution	224×224	224×224	224×224	224×224	224×224
Epochs	150	300	300	300	32
Batch Size	64	4096	1024	1024	32768
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Decay / $\beta_2$	1e-3 ( $\epsilon$ )	0.999	0.999	0.999	0.999
Momentum / $\beta_1$	0.9	0.9	0.9	0.9	0.9
Base Learning Rate	0.01	0.003	0.0005 × bs/512	0.001	5e-4
LR Schedule	OneCycle	Cosine decay	Cosine decay	Cosine decay	Cosine decay
Decay Rate	-	-	-	-	-
Warmup Epochs	0	5	5	20	-(2000 steps)
Warmup Schedule	-	-	-	Linear	-
Label Smoothing	-	0.1	0.1	0.1	-
RandAugment	(9, 0.5)	(9, 0.5)	(9, 0.5)	(9, 0.5)	(9, 0.5)
AutoAugment	-	-	-	-	-
Mixup	0.5	0.8	0.8	yes	yes
CutMix	0.5	1.0	1.0	yes	yes
Random Erasing $p$	0.25	0.25	0.25	0.25	yes
Dropout	-	0.1	0.1	-	-
Stochastic Depth	-	0.1	0.1	0.2	0.2
Drop Path	-	-	-	-	-
Layer-wise LR Decay	-	-	-	-	-
Weight Initialization	-	-	-	-	-
Layer Scale Init	-	-	-	-	-
Head Init Scale	-	-	-	-	-
EMA	-	-	-	-	-
Loss Function	CE	CE	CE	CE	InfoNCE (Contrastive)
Mixed Precision	Yes	Yes	Yes	Yes	Yes
Top-1 Accuracy	~82.0%	~83.0%	~83.0%	~83.0% (est.)	~78.0% (0-shot)

Table 11: Pretraining hyperparameter for models later fine-tuned on ImageNet-1K.

Model	Dataset	Epochs	Pretraining Setup
ConvNeXtV2-Tiny	ImageNet-22k	800–1600	AdamW, LR 1.5e-4, weight decay 0.05, $\beta_1=0.9$ , $\beta_2=0.95$ , Cosine decay, RandomResizedCrop, warmup 40 epochs.
EfficientNet-B5	JFT-300M	800	Noisy Student self-training with teacher on IN-1k, student on JFT-300M with RandAugment, Mixup, Dropout.

for Oxford-IIITPet. In the MI domain, training durations were 30 minutes for BloodMNIST, 3 hours for ChestMNIST and PathMNIST, 20 minutes for DermaMNIST, and 5 minutes for RetinaMNIST. Training on RS datasets took 90 minutes for AID, DeepGlobe, PatternNet, and RSD46-WHU, and 15 minutes for UCMerced.

Evaluation time per model and dataset ranged between 1 and 10 minutes, depending on the suppression condition and dataset size.

## L Ethics Statement and Risk Assessment

This study involved a low-risk visual classification task conducted with adult participants. All participants were volunteers, fully informed about the purpose and procedures of the study, and provided written informed consent prior to participation. No vulnerable populations (e.g., children, patients, or individuals with impaired consent capacity) were involved. Participants were not exposed to any physical or emotional risks, high stress levels, or invasive procedures such as fMRI or TMS.

The study design, including all procedures for participant interaction and data handling, was reviewed through the standard ethics assessment protocol of our institution. The responsible ethics committee certified that the study complies with all relevant legal and institutional guidelines. Specifically, the ethics committee confirmed that:



Table 12: Training hyperparameters for ResNet50 models across all domains and settings. RRC: RandomResizedCrop, HF: HorizontalFlip. All models were trained or fine-tuned using supervised learning with cross-entropy loss, except for ChestMNIST (binary classification) and DeepGlobe (multi-label classification), which used binary cross-entropy.

Dataset	Pretraining	Epochs	Batch Size	Image Size	Optimizer	LR	Weight Decay	LR Schedule	Train Augment
<i>Computer Vision (From Scratch)</i>									
ImageNet	–	100	256	224x224	SGD	0.1	0.0001	CosAnnealWR	RRC + HF
Flowers102	–	300	64	224x224	AdamW	0.001	0.01	StepLR (4)	RRC + HF
STL10	–	300	64	224x224	AdamW	0.001	0.01	StepLR (4)	RRC + HF
Caltech101	–	300	64	224x224	AdamW	0.001	0.01	StepLR (4)	RRC + HF
OxfordIIITPet	–	300	64	224x224	AdamW	0.001	0.01	StepLR (4)	RRC + HF
<i>Computer Vision (Pretrained)</i>									
Flowers102	IN-1k	100	64	224x224	AdamW	1e-5	0.001	CosAnnealWR	RRC + HF
STL10	IN-1k	100	64	224x224	AdamW	1e-5	0.001	CosAnnealWR	RRC + HF
OxfordIIITPet	IN-1k	100	64	224x224	AdamW	1e-5	0.001	CosAnnealWR	RRC + HF
Caltech101	IN-1k	100	64	224x224	AdamW	1e-5	0.001	CosAnnealWR	RRC + HF
<i>Medical Imaging</i>									
BloodMNIST	–	50	64	224x224	AdamW	0.001	1e-5	StepLR (3)	RRC + HF
ChestMNIST	–	50	64	224x224	AdamW	0.001	1e-5	StepLR (3)	RRC + HF
DermaMNIST	–	50	64	224x224	AdamW	0.001	1e-5	StepLR (3)	RRC + HF
PathMNIST	–	50	64	224x224	AdamW	0.001	1e-5	StepLR (3)	RRC + HF
RetinaMNIST	–	50	64	224x224	AdamW	0.001	1e-5	StepLR (3)	RRC + HF
<i>Remote Sensing</i>									
AID	–	80	64	600x600	AdamW	0.0005	0.01	CosAnnealWR	RRC + HF
DeepGlobe	–	80	64	256x256	AdamW	0.0005	0.01	CosAnnealWR	RRC + HF
PatternNet	–	80	64	256x256	AdamW	0.0005	0.01	CosAnnealWR	RRC + HF
RSD46-WHU	–	80	64	256x256	AdamW	0.0005	0.01	CosAnnealWR	RRC + HF
UCMerced	–	80	64	256x256	AdamW	0.0005	0.01	CosAnnealWR	RRC + HF

- the study involves no foreseeable risk of harm;
- participants are not drawn from vulnerable populations;
- data privacy is protected in accordance with applicable regulations;
- informed consent was obtained from all participants;
- the study team bears responsibility for the truthful completion of the ethics review questionnaire and the ethical integrity of the study.

As such, the study received approval from the institutional ethics committee, and no ethical concerns were identified.

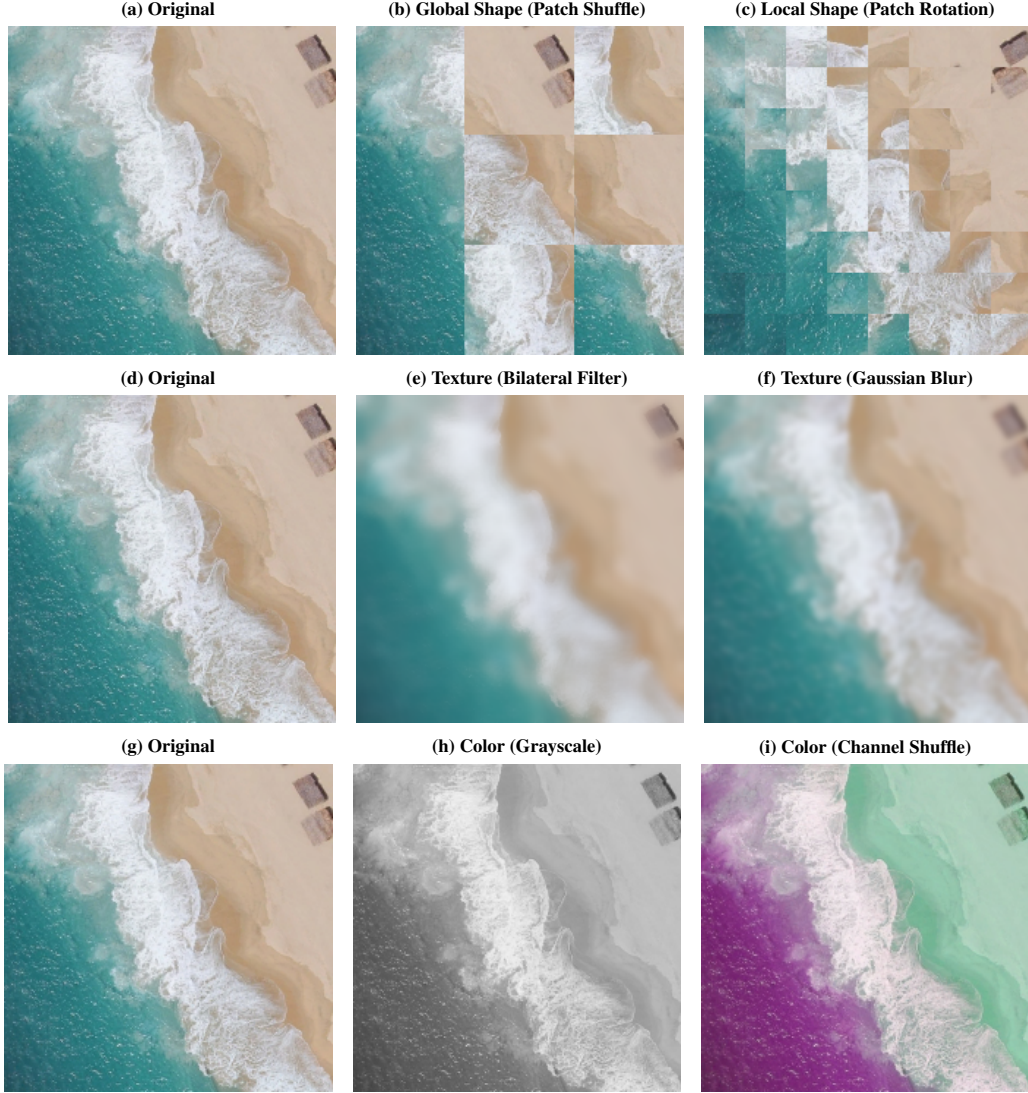


Figure 16: Visual illustration of feature suppression transformations applied to a sample image from the AID training set belonging to the class farmland. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.

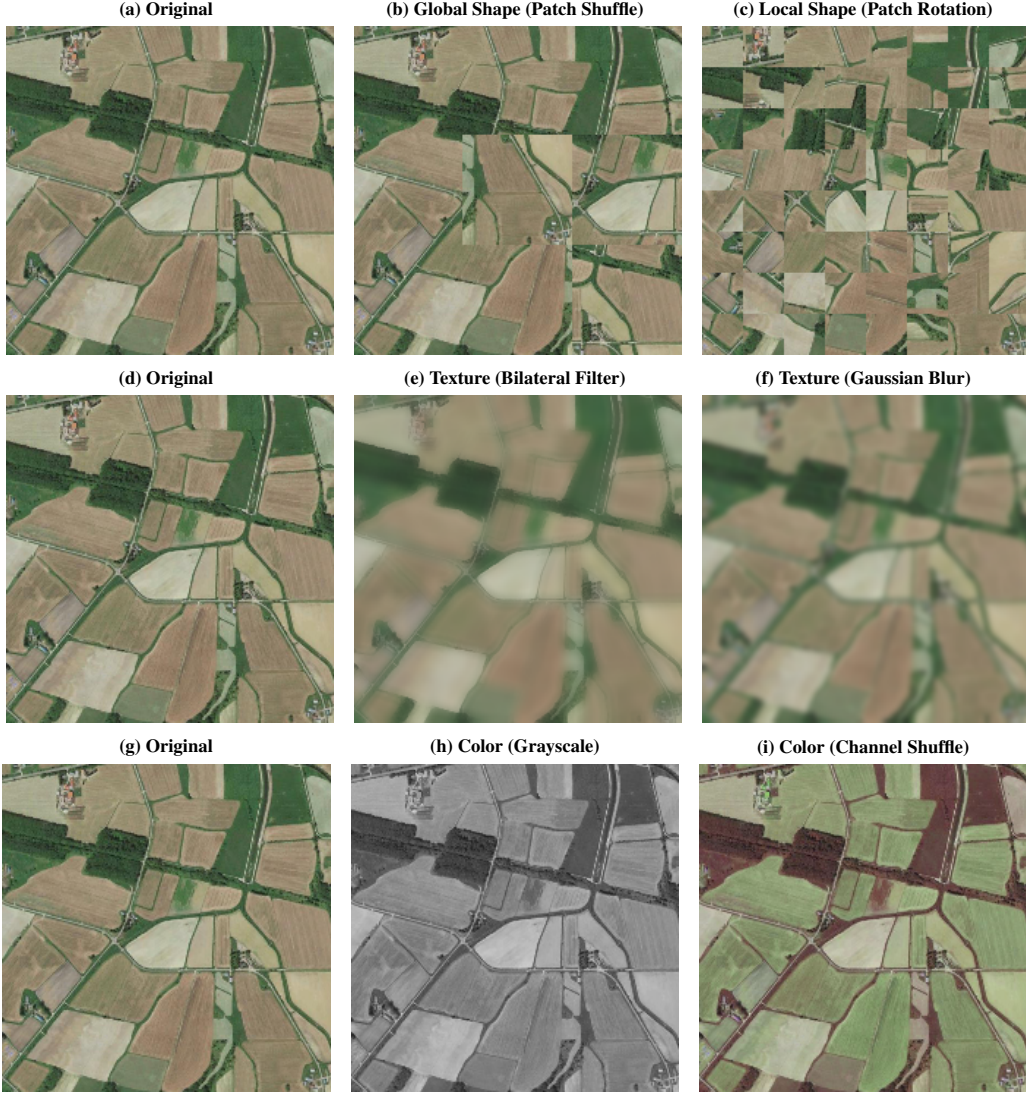


Figure 17: Visual illustration of feature suppression transformations applied to a sample image from the AID training set belonging to the class beach. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.

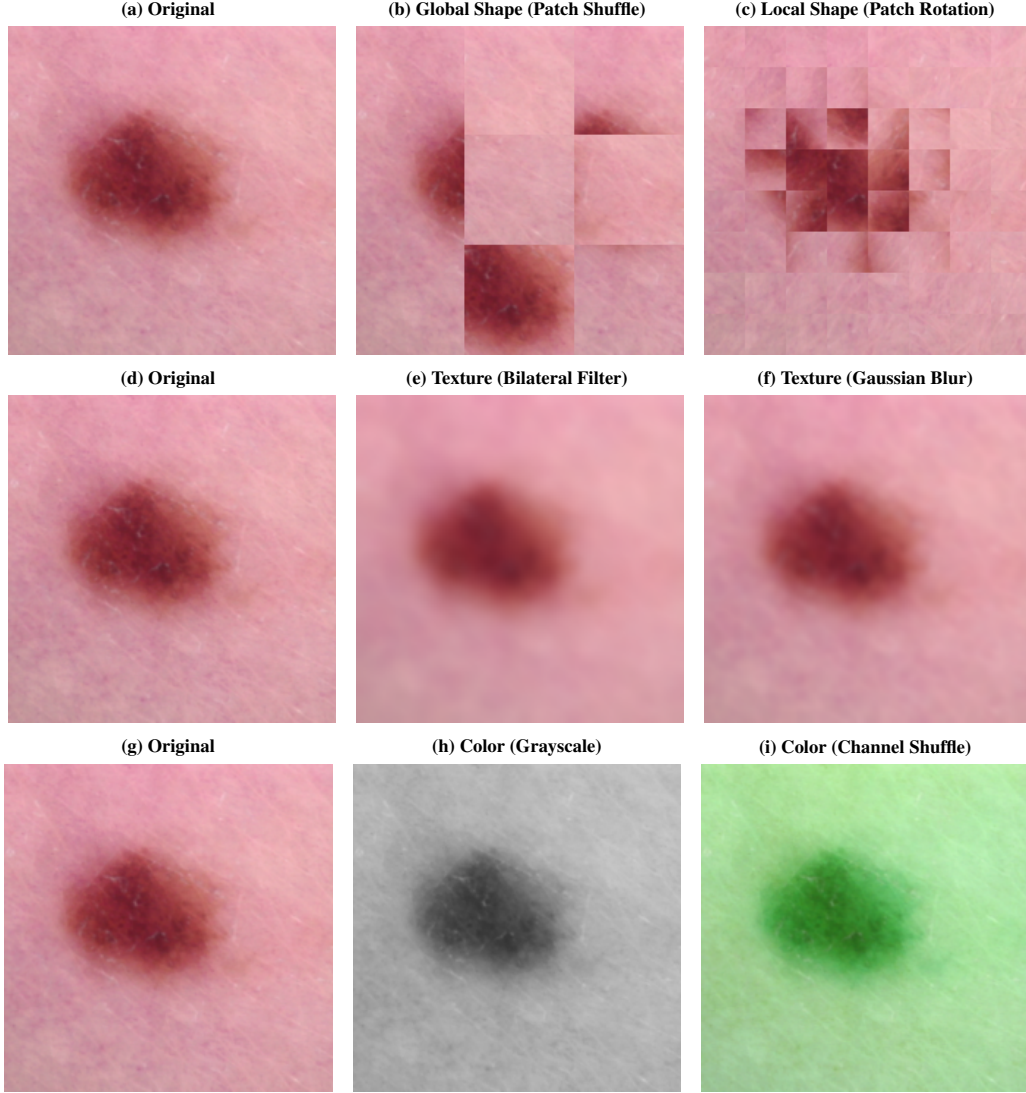


Figure 18: Visual illustration of feature suppression transformations applied to a sample image from the DermaMNIST training set belonging to class melanocytic nevi. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.



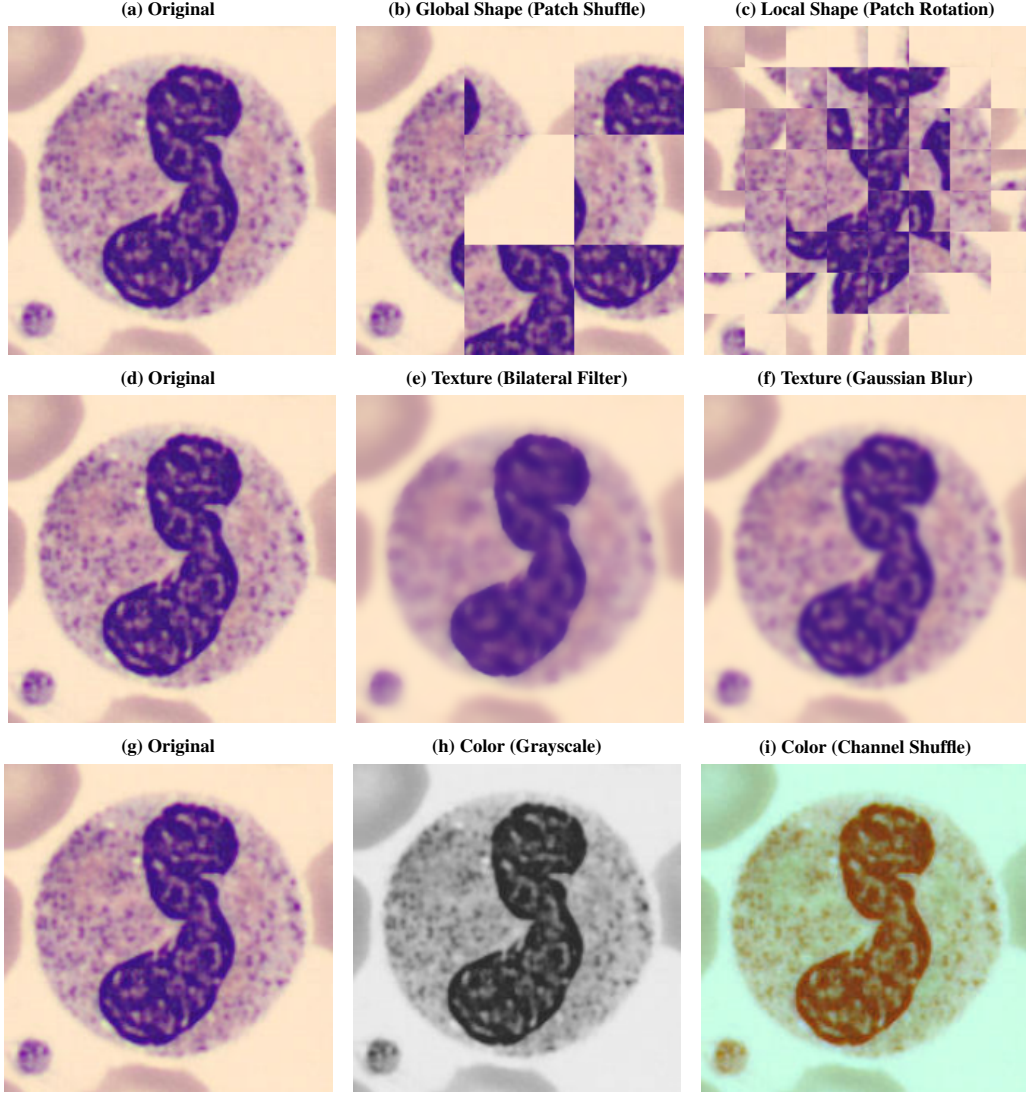


Figure 19: Visual illustration of feature suppression transformations applied to a sample image from the BloodMNIST train set belonging to the class neutrophil. **(a, d, g)** Show the original image. **(b)** Global shape suppression via Patch Shuffle with grid size 3. **(c)** Local shape suppression via Patch Rotation with grid size 8. **(e)** Texture suppression using Bilateral Filtering with  $d = 12$ ,  $\sigma_{\text{color}} = 170$ , and  $\sigma_{\text{space}} = 75$ . **(f)** Texture suppression using Gaussian Blur with kernel size  $k = 11$  and standard deviation  $\sigma = 2.0$ . **(h)** Color suppression via grayscale conversion. **(i)** Color suppression via random channel shuffle.