Advanced Topics in Machine Learning - PA1

Ahmad Ashraf [1] **Sarim Malik** [2]

Abstract

Different models in machine learning are preferred in different domains due to their inherent inductive biases. This paper explores these biases across three categories of models: discriminative (CNNs and ViTs), generative (VAEs and GANs), and contrastive (CLIP), and identifies their common use cases in real-world scenarios. Specifically, we conduct a thorough investigation into how each model's inductive bias impacts generalization. Our findings show that CNNs exhibit a strong texture bias due to the localized nature of convolutions, while ViTs and CLIP are more shape-oriented, owing to their architectural design. For generative models, we observe that VAEs produce lower-fidelity images, whereas GANs generate high-quality images at the cost of diversity, due to differences in their training objectives. Overall, our results provide a diagnosis of how inductive biases shape model behavior and inform their suitability for different applications. Our codebase with complete experiments can be found here: https://github.com/ATML-AshrafxSarim/PA1 and backup: https://github.com/s-malix21/atml-fall2025.

1. Introduction

Inductive bias constitutes the built-in priors and assumptions in machine learning models that determine how they learn new data distributions. A prevalent problem in ML is that different architectures, such as discriminative models like CNNs, ViTs, generative models like VAEs and GANs, and contrastive models like CLIP, all have different and distinct inductive biases, therefore in each architecture, the features (such as texture and shape) that are specifically focused on are different and how models perform on OOD data may also differ. hence, comes the problem of OOD generalization, which becomes a major challenge. When these architectures are presented with data distributions, domains and styles that they haven't seen before during their respective training phases, they then fail to make accurate predictions. Therefore this problem of unexpected inputs becomes a really significant part of not just training models but also understanding how they are generalizing on data from OOD distributions.

This research aims to empirically analyze and delve deeper into how primarily model architectures and different kinds of data, influences model biases. The key research questions that we aim to tackle through this study are as follows:

- How exactly do DL architectures encode semantic information in vision tasks?
- In what ways do architectural design choices affect resilience to distribution shifts and totally unseen/new input domains?
- What properties of a given model or what strategies during training, are best capable of supporting reliable performance in unseen data

The approach that we undertake in this study is a configuration of systematic experiments on discriminative, generative, and contrastive models that isolate the affects of each model-specific/internal factors as well as external environment or data-specific factors, in order to discover and analyze training strategies for robust ML systems.

2. Methodology

2.1. Discriminative Models

2.1.1. Model Architecture, Setup, Datasets

We have utilised a ResNet-50 model, which is a pretrained ImageNet model but has been finetuned on CIFAR-10, with a modified fully connected layer. We are also utilizing a vision transformer from the ViT-S/16 family, with 16x16 patches which has been pretrained and finetuned using the **timm** library. Even though Cifar-10 has 32x32 images by default, both models first ressized the inputs to 224x224 and used a typical ImageNet normalisation applied on all images. The training configuration was as follows, we used Adam Optimiser for 5 epochs using L.R of 1e-3 for ResNet-50 and the same optimiser with an L.R of 5e-4 and 3 epochs for ViT-S/16.

The primary dataset that we employ is CIFAR-10 which has 50,000 training and 10,000 test samples. for the domain

shift evaluation tasks we employ the PACS dataset through the **flwr** library, which has 4 domains (photo, art painting, cartoon, sketch). The usage of the PACS dataset was done by using three domains as source domains for training and using the unseen domain as the target domain for evaluation, we rotated this between two separate domains. We also constructed custom augmented datasets which included grayscale, stylized (to simulate PACS), translated, patch permuted and occluded iterations.

2.1.2. Color Dependency; Shape vs Texture Bias

A metric that we used here to determine color bias was by taking difference between the original and grayscale performance. This was made easier by measuring accuracy drop from the set baseline to quantify color resilience of both our discriminative models.

For our stylized dataset, we employed 8 different texture/style modification techniques(checkerboard, stripes, cartoon, oil painting, watercolor, sketch, mosaic, noise patterns). We also applied texture transfer techniques such as bilateral filtering and edge detection using the **cv2** library.

We also computed the **shape bias metric** as the percentage of correctly classified shapes against all total predictions, and the texture conflicted with the shape for this particular metric (high shape bias, low texture bias). We evaluated 500 samples per model for this experiment.

2.1.3. Translation Invariance and Spatial Robustness

Translated images were produced with shift levels ranging from: 3, 5, 8, 16, 32 pixels. Reflection padding was used here for withholding image content. A **consistency metric** was computed as the percentage of images getting the same predicted label before and after translation. Similarly an **invariance metric** detailed the classification accuracy on the translated images for each discriminative model.

In order to test spatial robustness, we applied a patch permutation to our images, by shuffling a 32x32 pixel patches across the image to disturb the global structrue of the image, For occlusion testing, we incoporated 56x56 squares at random locations on the images. We then measured accuracy degradation to assess the reliance on spatial arrangement vs local features in each image.

2.1.4. Feature Representation and Domain Generalization

For feature representation analysis, we extracted layer features from our ResNet and ViT through the average pool and forward features. We applied PCA for 50 components followed by t-SNE for visualization. Normal and stylized images were combined into a single dataset to assess cross-domain clustering. Moreover, we used Silhouette score, separation ratio to estimate the cross-domain coherence for each model.

For the domain generalization tasks, we finetuned our selected models on 3 PACS domains, while testing was conducted on the held out domain. In one experiment we used "sketch" domain as the primary test case for shape vs texture bias hypothesis. The metric used here was accuracy drop from the in-distribution performance, and this was our robustness indicator for the domain shift induced in this experiment.

2.2. Generative Models

Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) were the chosen generative models for comparison to research their own inductive biases.

2.2.1. Dataset Selection

CIFAR-10 was chosen as the dataset for comparison between VAEs and GANs due to its small input size (32 x 32), making it feasible for faster training amongst both models. Furthermore, the variety of features represented within the multiclass dataset was another factor in its selection, in order to produce variation in results due to the inductive biases in each model.

2.2.2. Model Architecture and Setup

Variational Autoencoder. The variation of the VAE to be trained was the VAE variant with KL annealing for effective training. The model consisted of a simple convolutional encoder and decoder with a latent space of 64 chosen due to the small feature set of CIFAR-10. For training, the VAE was trained for 50 epochs using the Adam optimizer with default β_1 and β_2 values along with a learning rate of 1e-3.

Contrary to vanilla VAE training, the loss function was modified to promote initial image reconstruction with the modification to the usual ELBO Loss defined as:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|x_i - \hat{x}_i\|^2}_{\text{Reconstruction Loss (MSE)}} + \beta \underbrace{D_{\text{KL}}(q_\phi(z|x_i)\,\|\,p(z))}_{\text{KL Divergence Loss}} \tag{1}$$

The β value was slowly increased from 0 to 1 after each round till epoch 40, in order to promote both image reconstruction and prior-posterior matching.

Generative Adversarial Network. The GAN used was a simple convolutional-based DCGAN with the latent dimension chosen to be equal to a 100 for the generator and

a convolutional based discriminator as well. The model architecture for both models consisted of convolutional layers (transpose convolutions for the generator), 2D batch normalization layers, and ReLU layers with the final output squeezed through a Tanh layer.

The models were initialized with the DCGAN style weight layer initialization, filling the convolutional and batch normalization layers with a mean of 0 and a standard deviation of 0.02 in order for training stability.

Both the generator and the discriminator were trained with the same learning rate of 2e-4 on the Adam Optimizer, and trained with equal updates (1:1 training ratio per batch) for the same 50 epochs, to produce a fairer comparison with the VAE.

A different setup from Vanilla GANs was the use of the non-saturating loss for the generator and label-smoothing, where real labels were smoothed to 0.9 in order for stable learning and preventing common issues such as mode collapse.

2.2.3. Reconstructions vs. Generations

In order to test the generative capabilities of both models after training, VAEs were tested on their capabilities to reconstruct sample images and GANs were tested on their ability to generate images by sampling from random latent vectors.

For the VAEs, images were sampled from the test dataset of CIFAR-10 to test its ability to reproduce the input image. Images were reconstructed by passing the set of test images through the encoder to the latent space, and from the latent space through to the decoder to achieve the final image. For the GAN, random latent vectors were sampled with the latent dimension of a 100 and passed through the generator.

2.2.4. Inception Score and Frechet Inception Distance (FID)

The Inception Score is a metric used to assess the quality and diversity of an image through a pre-trained Inception network. Samples from both VAEs and GANs were passed through the network to compare the quality and diversity of the images. Higher inception scores indicate more realistic (high fidelity) images and lower inception scores indicate blurry or less diverse images.

The Frechet Inception Distance (FID) is a metric used to assess the quality of images against a reference dataset, by comparing the feature distributions of images between the real and generated datasets. Lower FID scores indicate generated images are closer to real images in feature space. We compute an FID score for both VAE and GAN using a 1000 generated samples, and compare it against 1000 real samples of CIFAR-10 by passing them through the *inception-v3-compat* feature extractor in order to compare their similarity with real images.

2.2.5. Latent Interpolation and Latent Representation Analysis

Next, we explore the latent space properties of each model. For the VAE, two test images were sampled from the test dataset and mapped to latent space using the VAE encoder. A latent interpolation was then performed between the two latent vectors z_1 and z_2 and decoded through the VAE decoder. On the other hand, for the GAN, two random latent vectors were sampled from the GAN latent space. A latent interpolation was then performed between these vectors and passed through the generator.

Furthermore, latent representation analysis is performed on the VAE by performing t-SNE on the latent vectors of a sample of test images. This is done in order to investigate the structure of the latent space and assess the clustering capabilities of the VAE.

2.2.6. OOD Input Behavior

We lastly evaluate how robust our models are in handling samples outside their training input. We evaluate the VAE's performance based on reconstruction MSE loss when facing OOD data versus normal data. Since there is no direct measure for the GAN, we test the GAN's ability to generate samples when the scale of the normal distribution through which the latent vectors are produced is increased, changing the range from $\mathcal{N}(0, 1)$ to $\mathcal{N}(0, 10)$.

2.3. Contrastive Models

2.3.1. Experimental Setup

The experimental setup for this study utilizes a pretrained CLIP ViT-B/32 without fine-tuning, and the ResNet-50 architecture fine-tuned on CIFAR-10, utilized in the discriminative models' study. The datasets utilized include CIFAR-10, and the PACS dataset for the domain shift testing. We also utilize multiple text templates for the CLIP model. The primary classification protocol is cosine similarity between the image embeddings and the text embeddings for each class, additionally we also include the top-1 accuracy and the per-class accuracy for enhanced analysis.

2.3.2. Multimodal Representation Analysis

We propose the usage of a bidirectional matching for image to text retrieval using cosine similarity between embeddings. The feature space visualization utilizes a mixed dataset of 600 images from normal CIFAR, sketch style and grayscale images transformed. For the dimensionality reduction, we employ t-SNE to both CLIP and ResNet-50 embeddings

2.3.3. Shape vs Texture Bias Evaluation

We construct a cue-conflict dataset using 30 texture modified images that in turn have 6 shape-texture pairs (cat-elephant, dog-horse, etc.)

$$\text{Shape Bias} = \frac{\text{Shape-consistent predictions}}{\text{Total predictions}} \times 100\%$$

For the baseline comparison, an identical evaluation was applied to ResNet-50 using classification predictions.

2.3.4. Robustness Assessment

The corruption types we employed consisted of the following: Gaussian noise ($\sigma = 0.1$), Gaussian blur ($\sigma = 2.0$), high contrast ($3\times$ enhancement). We used 50-image subsets per corruption type.

Degradation Metric:

$$\Delta\text{acc} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{corrupted}}$$

The cross model comparison conducted here was between the CLIP zero-shot and the supervised ResNet-50.

3. Results

3.1. Discriminative Models

3.1.1. In-Distribution and Basic Robustness Results

(This section includes baseline accuracy, color bias, translation invariance, permutation, and occlusion tests)

Model	Shape Bias (%)	Grayscale Acc. (%)	Accuracy Drop (%)
ResNet-50	36.00	15.80	73.31
ViT	60.60	35.90	54.38

Table 1. Comparison of ResNet-50 and ViT biases under shape and color-based evaluations.

For the baseline in-distribution performance, the ViT achieves an accuracy of 0.9028 while the ResNet achieves a competitive accuracy of 0.8911 on the CIFAR-10 test set. The modest difference can be attributed to the different training configurations used for each model's training.

ResNet-50 shows severe color dependency with 73.31pc accuracy drop on grayscale images. Whereas ViT demonstrates superior color robustness with only 54.38pc accuracy drop. It can be inferred that with regards to **color bias**, ResNet's local receptive fields emphasize texture and color patterns over global shape information (this will be confirmed by the ablation on its texture bias). Whereas ViT's attention mechanism enables shape-based detection. The ViT also exhibits stronger shape bias (60.60pc) compared to ResNet-50 (36.00pc). ViT has teh global context during the computation of its self attention (X matrix gets multiplied as a whole, so no reduced receptive field), in otehr words ViT's shape bias comes from patch-based processing and global attention. A key observation from this experiment is that models that rely more on shape can show a better performance when color information is removed, as observed in our metrics.

For the translational invariance, we observe that ResNet shows slightly higher consistency at all shifts (98.00pc vs 96.83pc at 3px), this happens due to the translational equivariance that is inherent to convolution for translational shifts. The pooling operations in ResNet50 help with reducing reliance on spatial sensitivity.

it has to be noted here that both model performances degrade with larger shifts, since at that point objects may move significantly out of frame, thus causing teh degradation

For the patch permutations the ViT significantly outperforms the ResNet(48.80pc vs 19.20pc), whereas the same phenomena is observed for the occlusion tests (89.80pc vs 81.00pc). For both cases, the ResNet struggles more because its receptive fields are more susceptible to disturbances, as we already established that it depends more on spatial information. Whereas global attention of a ViT can leverage information from non-occluded patches, hence maintaining performance even when some patches are unavailable.

Test / Shift	**ResNet-50 (%)**		**ViT (%)**	
	Consistency	Accuracy	Consistency	Accuracy
Patch Permutation	–	19.20	–	48.80
Occlusion	–	81.00	–	89.80
Shift = 3	98.00	88.33	96.83	90.17
Shift = 5	96.33	91.17	95.00	90.33
Shift = 8	95.83	88.67	92.00	88.50
Shift = 16	94.67	88.50	93.00	88.67
Shift = 32	92.17	84.00	91.67	87.67

Table 2. Robustness of ResNet-50 and ViT under patch permutation, occlusion, and translation shifts.

3.1.2. Feature Representation Analysis; Domain Generalization Performance

(This section includes contains t-SNE/PCA feature space analysis, PACS cross-domain results, leave-one-domain-out experiments, and robustness insights)

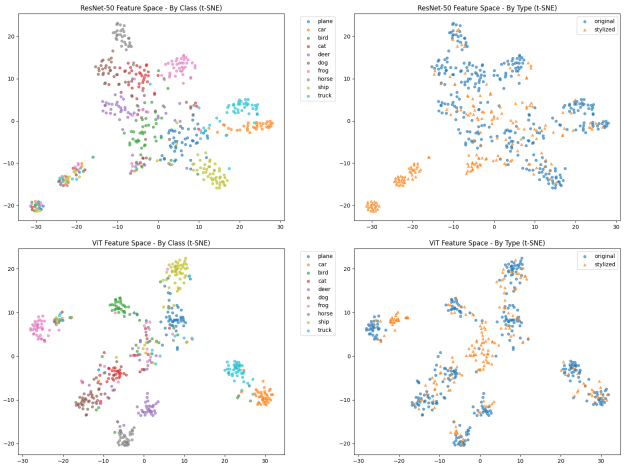

Figure 1. Visualization of Feature Space (by Type+Class

Metric	ResNet-50	ViT
Class Silhouette Score (↑)	0.049	0.173
Average intra-class distance (↓)	14.195	10.435
Average inter-class distance (↑)	22.661	24.595
Separation ratio (↑)	1.596	2.357
Stylized → Original distance (↓)	12.530	6.384
Average neighborhood density (↓)	1.156	0.954

Table 3. Feature space analysis of ResNet-50 and ViT. Arrows (↑/↓) denote whether higher or lower values are better.

In terms of the feature quality, we extended to an additional experiment apart from visualizing t-SNE plots and visualzie class based seperation metrics as well. It is observed that ResNet's feature space groups images with similar textures together, whereas ViT shows the better class seperation as well as a better distinction between original and stylized images, thus reflecting a better semantic focus for the ViTs. We supplement our visualizations with additional metrics. It is thus shown that ViT achieves superior class separation with 3.5x higher silhouette score (0.173 vs 0.049), and it also has a better intra-class clustering with 26pc lower intra-class distance (10.435 vs 14.195). This leads to a better seperation ratio for the ViT, due to the combined effect of the two metrics above. We attribute this superiority to the global attention mechanism of ViTs that allow it to better focus on class relevant fetaures, mitigate variations as opposed to Resnet who's feature aggregation, and its local receptive fields make it miss out on global discriminative features.

The PACS domain generalization results present a more complex picture than initially hypothesized. On the Sketch domain, ResNet unexpectedly outperforms ViT with an accuracy of 54pc to ViT's 37pc. The Art Painting domain shows nearly equivalent performance where for ResNet: 41.70pc, ViT: 41.41pc, suggesting that moderate stylistic variations affect both architectures similarly. This is contrasted to the fact that when we tested ViT and ResNet in a PACS simulated dataset using a transformed CIFAR-10 subset, we got higher acc values for the ViT across all test domains. This would imply that even though training dynamics differ between our architectures, since we had sam enumber of epochs (3) and the same optimiser for finetuning our two disc models on the domain shifted dataset, that could be a potential factor. ResNet may adapt faster to new domains during fine-tuning despite the worse initial inductive biases, while ViT's superior representational quality doesn't guarantee better domain transfer in as many epochs.

This additional experiment also proves something else. PACS domains differ in lighting, drawing styles, color distributions, and edge characteristics that make it far more complex in terms of its visual statistics. It may be the case that the hierarchical processing of ResNet provides better foundation for domain adaptation on new datasets like PACS when the training dynamics are not adjusted to cater for the specific requirements of each classifier.

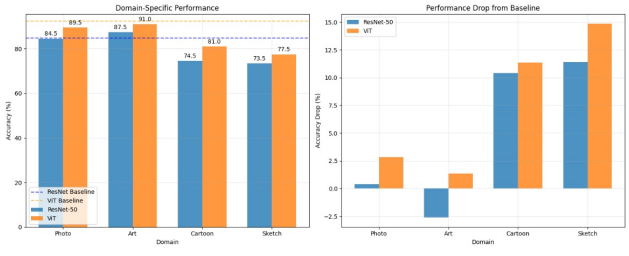

Figure 2. Accuracy comparison of ResNet-50 and ViT across simulated PACS domains using CIFAR-10

Held-out Domain	ResNet Accuracy (%)	ViT Accuracy (%)
Sketch	54.11	37.57
Art Painting	41.70	41.41

Table 4. Domain generalization results on the PACS dataset.

3.2. Generative Models

3.2.1. Reconstructions vs. Generations

Figure 3 highlights the different capabilities of VAEs and GANs in reconstruction and generation. VAE reconstructions tend to be blurrier versions of the original image (missing certain edges and details), while GAN outputs are generally sharper and capture finer details but they tend to introduce small artifacts that make the images look slightly odd. However, through qualitiative analysis, GANs exhibit lower diversity when it comes to generating samples.

Figure 3. Original Image vs. VAE Reconstruction vs. GAN Generation

3.2.2. Inception Score and Frechet Inception Distance

To quantify the image quality deductions made in 3.2.1, we rely on Inception scores to assess the quality and diversity of the generated images. As shown in Table 5, the VAE attains a lower inception score than the GAN on the same sample size, confirming the hypothesis that GANs produce high-fidelity samples, whereas VAEs tend to generate images that are less realistic in terms of quality.

Model	Inception Score
VAE	3.25 ± 0.18
GAN	5.53 ± 0.55

Table 5. Comparison of Inception Scores between VAE and GAN. Higher inception scores are indicative of high fidelity images.

Regardless of image quality, we also evaluate how closely the generated images resemble real CIFAR-10 samples by computing the Fréchet Inception Distance (FID) on a fixed number of samples. As shown in Table 6, the GAN achieves a substantially lower FID score (nearly half that of the VAE) indicating that GANs generate images more similar to the real dataset. In contrast, VAEs tend to produce blurrier, less realistic images due to their reliance on minimizing the MSE loss, which encourages averaging effects.

Model	Frechet Inception Distance
VAE	164.32
GAN	86.98

Table 6. Comparison of FID scores between VAE and GAN. Lower FID scores indicate higher similarity between generated and real images.

3.2.3. Latent Interpolation and Latent Representation Analysis

Latent interpolation sequences are used to compare the properties of the latent spaces of the VAE and GAN. Figure 5 illustrates the differences in interpolation behavior between the two models. The VAE produces smooth transitions of features when moving from one latent vector to another. In contrast, while the GAN also attempts to achieve smooth interpolation, it often does so at the cost of image quality, introducing abrupt random artifacts.

This can be explained by the fact that VAEs are trained with a KL-divergence term (Equation 1), which encourages the latent space to follow a smooth Gaussian distribution, leading to more continuous interpolations. In contrast, GANs learn a direct mapping from latent vectors to images without such regularization, often resulting in less coherent interpolations.

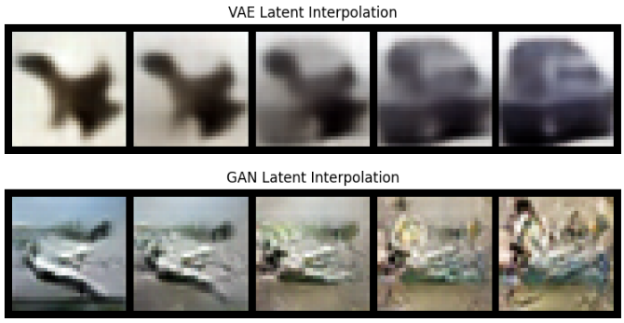

Figure 4. A sample latent interpolation sequence of GAN and VAE.

The latent space can be further examined for VAEs (unlike GANs, whose latent space is largely unstructured) by performing t-SNE on the latent vectors of a sample of test images. Interestingly, our VAE did not exhibit clear clustering of classes, shown by Figure 5. One reason is that the model was trained with KL annealing, which allowed it to prioritize reconstruction during early training. As a result, the posterior remained relatively weak in its ability to impose structure on the latent space.

By contrast, using a pure β-VAE with $\beta > 1$ would encourage a stronger disentanglement of features, which would possibly lead to a more structured latent space.

A latent traversal was performed to examine whether certain dimensions of the latent space captured interpretable factors. While most dimensions (see Appendix A.2.2) did not show clear semantic meaning, one latent dimension appeared to control the color of the main object (ranging from white to black). This suggests that using a β-VAE could further improve feature disentanglement.

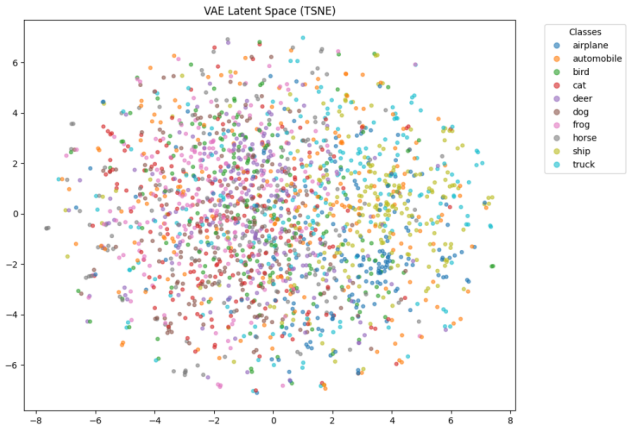

Figure 5. t-SNE representation of a 1000 latent vectors sampled from the VAE encoder.

3.2.4. Domain Generalization Behavior

On OOD data, specifically CIFAR-100 images, VAEs tend to show higher reconstruction error per image, which is expected. However, the difference in results is relatively small. A deeper investigation of the feature spaces of both CIFAR-10 and CIFAR-100 using t-SNE (Appendix A.2.1) reveals substantial overlap, suggesting that anomaly detection based solely on reconstruction MSE as a detection signal depends on factors beyond reconstruction quality and is heavily influenced by the similarity of feature distributions across datasets.

Table 7. Mean reconstruction error per image for VAE on in-distribution (CIFAR-10) and OOD (CIFAR-100) datasets.

Dataset	Mean Reconstruction Error
CIFAR-10	0.0368
CIFAR-100 (OOD)	0.0382

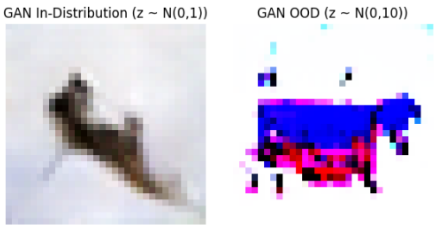

Figure 6. GAN generation from a latent vector sampled from $\mathcal{N}(0,1)$ compared to the same vector scaled to $\mathcal{N}(0,10)$.

For GANs, OOD samples result in extremely poor generations, with the images often being completely unrecognizable. This shows that GANs mainly learn a direct mapping from the latent space to image space and struggle to generalize beyond it.

3.2.5. Training Dynamics

VAE. The VAE was much simpler to train than the GAN because of its clear training objective. However, a tradeoff was evident between reconstruction quality and posterior matching. To balance this, KL annealing was applied, allowing the model to gradually learn both reconstruction and latent regularization.

GAN. The GAN was more difficult to train, often suffering from instability when using standard weight initialization. Techniques such as label smoothing and careful tuning of learning rates were necessary to stabilize training. In some runs, mode collapse occurred, showing that while the GAN could produce higher-fidelity images, it did so at the cost of diversity, failing to represent all training classes. Ultimately, using DCGAN weight initialization and the above stabilizing techniques led to a more stable minimax dynamic.

Prompt Template	Accuracy (%)
`a sketch of a {}`	85.14
`a cartoon drawing of a {}`	85.21
`a painting of a {}`	86.18
`a photo of a {}`	84.96
`a line drawing of a {}`	85.52
`an outline of a {}`	83.38

Table 8. Zero-shot classification accuracy on PACS (sketch domain) across different prompt templates

.

3.3. Contrastive Models

3.3.1. Zero-Shot and Cross-Domain Classification

Zero-shot classification of CIFAR-10 using CLIP produced surprisingly strong results, reaching approximately **87% accuracy**. This highlights the strength of CLIP's training and semantic alignment, especially when compared to the fine-tuned baselines in Task 1, which were trained specifically for this dataset and achieved around 90% accuracy.

Furthermore, on domain-shifted classification, Table 8 highlights the robustness of CLIP in cross-domain tasks. When compared to ResNet-50's performance on the sketch dataset, the gap is staggering (54.11% accuracy for ResNet vs. nearly 80% for CLIP).

This result aligns with our later experiments, where we observe CLIP's strong shape bias, which, combined with its extensive pretraining, allows it to outperform ResNet,

whose CNN backbone is more texture-biased and less robust to such domain shifts.

3.3.2. Image-Text Retrieval

CLIP demonstrates strong multi-modal alignment in the image-to-text retrieval task, where we extract the top-1 text match for a small set of images with fixed prompts. As shown in Table 9, the retrieved class is correct in most cases. Even in the few cases where CLIP fails, the retrieved text is often semantically close to the true label in the embedding space (e.g., 'automobile' vs. 'truck'), reflecting the strength of its learned representations.

Image ID	Ground Truth	Retrieved Text (sim)
0	truck	'a photo of a truck' (0.268)
1	truck	'a photo of a automobile' (0.274)
2	airplane	'a photo of a airplane' (0.261)
3	bird	'a photo of a bird' (0.276)
4	ship	'a photo of a ship' (0.273)
5	cat	'a photo of a cat' (0.274)
6	horse	'a photo of a horse' (0.302)
7	bird	'a photo of a bird' (0.315)
8	automobile	'a photo of a automobile' (0.283)
9	automobile	'a photo of a automobile' (0.272)
10	truck	'a photo of a truck' (0.285)
11	cat	'a photo of a bird' (0.258)
12	truck	'a photo of a truck' (0.308)
13	frog	'a photo of a frog' (0.274)
14	frog	'a photo of a frog' (0.231)
15	frog	'a photo of a frog' (0.281)
16	airplane	'a photo of a airplane' (0.252)
17	cat	'a photo of a cat' (0.265)
18	cat	'a photo of a dog' (0.271)
19	bird	'a photo of a bird' (0.298)

Table 9. Image-to-Text retrieval results for Task 3.3. Similarity scores (sim) indicate the model's confidence.

3.3.3. Prompt Engineering

Table 10 highlights the effect of different prompting strategies and indicate how presence of context helps CLIP in terms of classification accuracy (especially when comparing `{class}` vs. `a photo of {class}`).

Prompt Template	Accuracy (%)
`a photo of a {}`	87.80
`a picture of a {}`	87.95
`{}`	85.02
`a picture of a {} on grass`	87.37

Table 10. Zero-shot classification results for CLIP on CIFAR-10, with accuracies under different prompt templates.

3.3.4. Representation and Bias Analysis

Cross-Domain Representation Analysis. Approximately 1,500 augmented CIFAR-10 samples, containing an equal mix of normal, greyscale, and manually created sketch-based images, were passed through the feature extractors of CLIP and ResNet-50, after which t-SNE was performed (images shown in Appendix A.3). A critical issue in the process was the transformations applied, particularly for the sketch dataset, which slightly disturbed the feature space. Surprisingly, both ResNet and CLIP were relatively able to maintain their clusters. Cross-domain similarity was then computed to provide a quantitative measure, with CLIP slightly outperforming ResNet, achieving a cross-domain similarity of 0.771 against ResNet with 0.765. CLIP's performance could have been higher if the domain differences were more pronounced; ResNet was able to decipher greyscale images, but the sketch domain did not produce sufficiently good data to be useful.

Shape vs. Texture Bias. Table 11 highlights and reaffirms CLIP's strong shape bias. Across 30 cases, CLIP consistently favors shape-based cues. In the few cases where it instead relies on texture, this often occurs because the compared classes already share a high degree of shape similarity (e.g., dogs and horses are both four-legged animals). On the other hand, the same analysis is performed for ResNet-50 which exhibits a clear texture bias (results shown in Table 12).

Conflict Type	Shape Bias (%)	Texture Bias (%)
Cat vs. Elephant	40.0	60.0
Dog vs. Horse	60.0	40.0
Bird vs. Airplane	80.0	20.0
Ship vs. Truck	80.0	20.0
Horse vs. Deer	100.0	0.0
Frog vs. Truck	100.0	0.0
Overall (30 cases)	**76.7**	**23.3**

Table 11. Comprehensive shape vs. texture bias analysis for CLIP on 30 cue-conflict examples. Results show a strong preference for **shape** (76.7%) over texture (23.3%).

Decision Type	Count	Percentage
Shape-biased	5	16.7%
Texture-biased	25	83.3%

Table 12. Comprehensive shape vs. texture bias analysis for ResNet-50 on 30 cue-conflict examples.

Robustness Analysis. CLIP remains reasonably robust under different input perturbations, despite not being trained in adversarial settings. The observed drop in accuracy is

expected: for example, blur and high contrast can distort or obscure the shape structure, while noise disrupts fine-grained details. Regardless, CLIP still maintains relatively strong performance across these shifts, highlighting its robustness compared to standard supervised models.

Condition	Accuracy	Degradation
Original (CIFAR-10)	0.8795	–
Noise	0.7600	0.1195
Blur	0.8000	0.0795
High Contrast	0.8000	0.0795

Table 13. Robustness comparison of CLIP under different perturbations. Accuracy drops relative to the original CIFAR-10 evaluation.

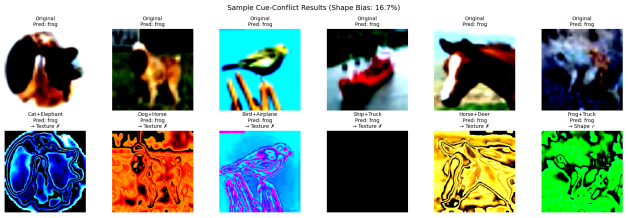

Figure 7. Visualizing sample cue-conflict examples

3.3.5. Interpretation of CLIP Biases

While CLIP is based on a ViT architecture, it shows an inherent inclination towards shape bias. A possible explanation is that, in order for the text and image embeddings to align as closely as possible, the model had to learn features that make it more robust. Rather than relying on superficial indicators (e.g., a CNN recognizing the background of water to classify fish), CLIP is forced to focus on cues such as the actual shape of the object. In contrast, discriminative models require explicit mechanisms like regularization or data augmentation to move beyond surface-level cues. This is consistent with our robustness experiments (Table 13), where CLIP maintains stronger performance under perturbations that distort texture but preserve shape.

4. Discussion

4.1. CNNs vs. ViT Bias

The results confirm CNNs' texture bias versus ViTs' shape bias. ResNet-50's 73.31pc accuracy drop on grayscale images compared to ViT's 54.38pc drop demonstrates heavy color dependence in CNNs. ViT's superior shape bias, 60.60pc against the 36.00pc, reflects its global attention mechanism enabling holistic object recognition, while ResNet's local receptive fields emphasize texture patterns. ViT's 3.5x higher silhouette score confirms better semantic feature organization over ResNet's texture-based representations.

4.2. Translation Invariance and Spatial Processing.

ResNet showed better translation consistency across multiple pixel shifts, confirming convolutional weight sharing provides natural translational equivariance. However, ViT's superior robustness to patch permutations which was 49pc agianst 19pc, and occlusions which was 90pc against 81pc; demonstrates attention mechanisms' flexibility with spatial disruptions. While CNNs excel at local transformations they were designed for, ViTs adapt better to structural noise/obstructions that requiring global context.

4.3. CLIP's Semantic Inductive Bias.

CLIP's higher shape bias (77pc) and cross-domain robustness (85pc on sketch vs ResNet's 54pc) show that multimodal training of vision-language models develop more structured feature hierarchies than purely visual architectures.

4.4. Fidelity-Diversity Tradeoff

The fidelity-diversity trade-off between VAEs and GANs was clearly demonstrated by our experiments highlighting each models generation capabilities. VAEs are trained with an objective of both image reconstruction and prior-posterior matching, showcasing a tradeoff. In the pursuit to maximize both of them, while maximizing diversity, the inception Score (3.25) and FID (164.32) were both significantly worse than the GANs, confirming the lower quality of their generations. In contrast, GANs generated sharp and high quality images (Inception Score of 5.53; FID of 86.98) by learning a direct mapping from a latent vector to an image. However, this came at the cost of diversity, as GAN training inherently does not cover class diversity; it just requires enough data to fool the discriminator. This makes GAN prone to a lot of training instability and requires experimental changes before leading to an ideal model.

4.5. Generalization

In terms of generalization, models with shape bias and semantic understanding performed better on new, unseen data. ViT's focus on object shapes and CLIP's text-image training helped them handle different visual styles, while ResNet's texture focus made it struggle with stylistic changes. However, these advantages require more training data, for eg ViTs need large datasets and more complex training dynamics (in terms of optimisers, weight decay and epochs) to learn spatial relationships that CNNs get naturally, and CLIP needs diverse text-image pairs.

4.5.1. Critical Reflection and Limitations

Several unexpected results warrant discussion. ResNet's superior performance on PACS sketch domain contradicts our shape bias hypothesis, suggesting that domain adaptation dynamics may favor hierarchical feature processing over attention-based mechanisms when training epochs are limited. The modest difference in VAE reconstruction error between CIFAR-10 and CIFAR-100 which was 0.0368 against 0.0382, indicates that reconstruction-based OOD detection is less reliable than expected, possibly due to feature overlap between natural image datasets.

5. Conclusion

Our analysis reveals that inductive biases are the key driver of model behavior and generalization. CNNs rely on texture due to local processing, while ViTs develop shape based representations through global attention. Generative models show a clear fidelity-diversity trade-off - GANs produce sharp but unstable outputs, while VAEs generate diverse, structured samples. CLIP's multimodal training creates human-like semantic understanding that transfers well across domains.

The critical finding is that human-aligned biases lead to better generalization. Models with shape bias (ViTs), semantic alignment (CLIP), or structured priors (-VAEs) handle distribution shifts much better than those relying on surface statistics. Models without these structured biases may work well in-domain but struggle when conditions change.

Future work could focus on bias-aware fine-tuning strategies that gradually introduce texture variations during training may help models develop robust shape representations without requiring architectural changes.

References

A. Appendix

A.1. Supplementary Material - Descriminative Models

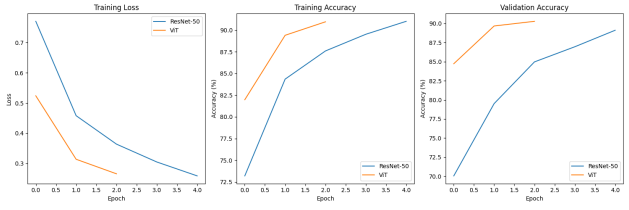

Figure 8. Report: Training Dynamics of Discriminative Models during finetuning on CIFAR-10

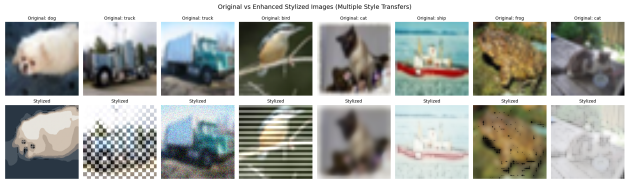

Figure 9. Visualization of stylized images using advanced texture modifications including cartoon style transfer

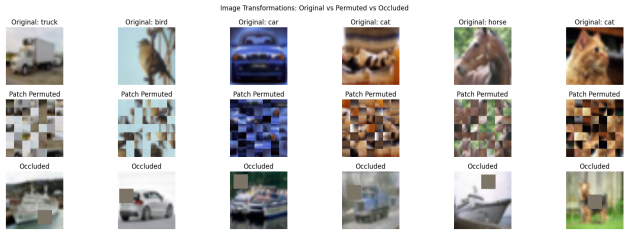

Figure 10. Visualization of Image Transformations: Original vs Permuted vs Occluded

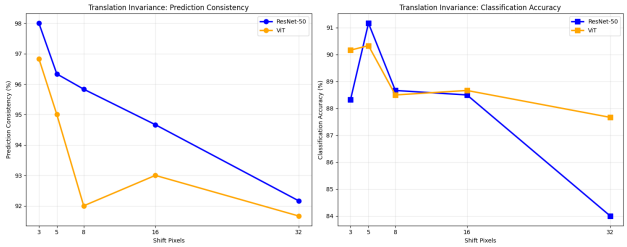

Figure 11. Visualization of TRANSLATION INVARIANCE ANALYSIS

A.2. Supplementary Material - Generative Models

A.2.1. CIFAR-10 vs. CIFAR-100 Feature Space Analysis

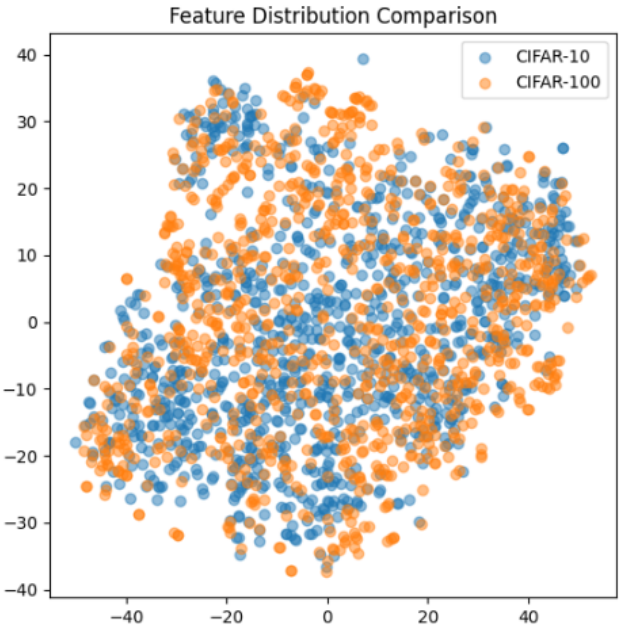

Figure 12. Feature Space Analysis of CIFAR-10 and CIFAR-100 using t-SNE and pretrained ResNet18's feature extractor.

A.2.2. Semantic Factor Investigation

Figure 13. Latent traversal within 10 dimensions for a sample image.

A.3. Supplementary Material - Contrastive Models

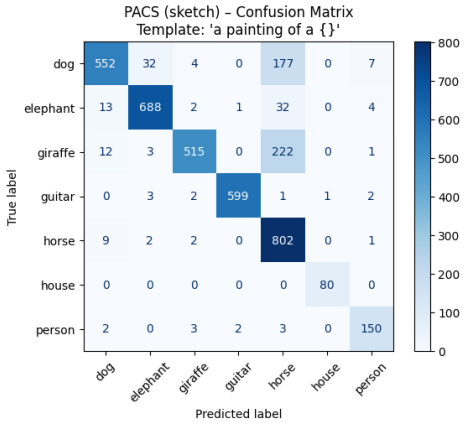

Figure 14. PACS (Sketch); Confusion Matrix for Domain-Shifted Classification

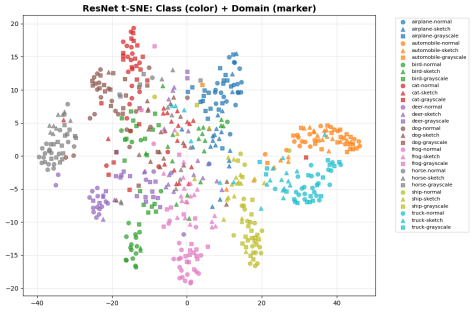

Figure 15. ResNet-50 t-SNE projection.

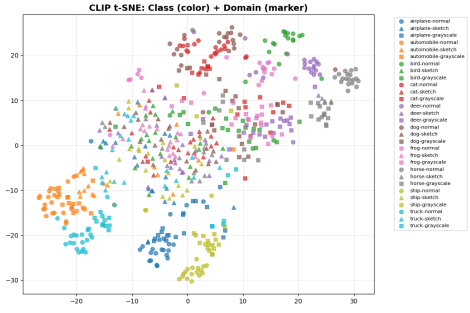

Figure 16. CLIP t-SNE projection.