How Inductive Biases Impact Generalization in Deep Vision Models

Qasim Ayub [1] **Usman Ahad** [1] **Muhammad Usman** [1]

Abstract

Deep learning vision models encode different inductive biases through their architectures, training objectives, and data sources. These biases determine what cues models prioritize (e.g., texture, shape, or semantics), how they represent information, and how well they generalize to out-of-distribution (OOD) data. In this paper, we conduct hypothesis-driven experiments to compare inductive biases across three families of models: discriminative (ResNet-50 vs. ViT-S/16), generative (VAE vs. GAN), and contrastive multimodal (CLIP ViT-B/32). The GitHub repository link for this project can be found here .

1. Introduction

Although deep learning models can learn hidden and intricate features of any dataset that they come across, they can sometimes focus on the wrong things or learn something too specific to the dataset itself. These models come with their own built-in assumptions called Inductive Biases that shape how they learn and generalize. CNNs, for example, rely heavily on locality and translation invariance, while Vision Transformers (ViTs) drop most of those assumptions in favor of global self-attention. On the other hand, multimodal models like CLIP inherit their biases not just from architecture but also from the training objective of aligning images and text at scale. These differences end up playing a central role not only in how models succeed on their training data, but also in how they behave under distribution shifts, where robustness is really tested.

In this project, we explore these biases across three categories of models: discriminative (ResNet-50 vs. ViT-S/16), generative (VAE vs. GAN), and multimodal contrastive (CLIP ViT-B/32). Each task is designed to bring out how different choices, i.e, convolution vs. attention, reconstruction vs. adversarial training, or vision-language alignment, shape what the model learns, what invariances it builds in, and how it handles out-of-distribution (OOD) inputs.

- **Task 1** compares CNNs and ViTs on classification under distribution shifts, highlighting their reliance on texture versus shape and testing their architectural invariances.
- **Task 2** looks at generative models, contrasting the VAE's smooth but blurry reconstructions with the GAN's sharp but sometimes unstable generations, and linking these outcomes back to their objectives.
- **Task 3** investigates CLIP, showing how multimodal supervision biases it toward more semantic, shape-oriented representations that transfer well in zero-shot and cross-domain settings.

The overall aim is to answer a simple but central question: how do architecture, training objectives, and data together shape the inductive biases of vision models, and what does that mean for representation, generation, and generalization beyond the training distribution? The following sections are structured around hypothesis-driven experiments, with a focus on analysis and insight rather than exhaustive training.

2. Discriminative Models (ViT vs ResNet)

2.1. Methodology

In this section we will try to analyse the performance and internal biases of ViT-S/16 and ResNet50. Both models will be pretrained on imagenet and we will fine tune the classification head for 2-3 epochs to achieve over 90% base test accuracy. All models will use Cross Entropy Loss and Adam optimizer with a learning rate of 0.001 for training. We will use STL-10 dataset for all tests to analyse differences in the underlying implementation of both models. For all datasets in this analysis we will be normalising our images according to imagenet mean and standard deviation. To analyse semantic biases like Color Bias, Shape Bias and Texture Bias, we will prepare 2 additional datasets using STL-10 test split; Gray scale dataset and Cue conflicted dataset, to see any performance changes after removing texture information from our images. While producing a gray scale dataset is reletively

[1]Department of Computer Science, LUMS, Lahore, Pakistan. Correspondence to: Qasim Ayub <27100168@lums.edu.pk>, Usman Ahad <27100041@lums.edu.pk>, Muhammad Usman <27100046@lums.edu.pk>.

easier, the cue conflicted dataset can be made using multiple approaches. We will be using a pretrained autoencoder fine tuned on VGG11 weights to merge each image of our dataset with a randomly selected image from a set of stylised images.

After analysing semantic biases we will test locality biases of our models. For this we will generate 3 additional test sets. First will be translated dataset which we will produce by translating each image by 40 pixels in a randomly chosen direction. Second will be shuffled dataset, where we will disrupt global structure of image by shuffling its patches. In shuffled dataset, we will try to maintain the local context of the images by shuffling patches of size 16x16. Third will be occlusion dataset where we will mask out patches of our images to make a few regions of our image unavailable. For this we will be masking 25 randomly selected patches of size 28x28 from our image.

Finally we will be analysing the domain generalisation of both models. For this test we will use PACS dataset. We will train ViT and Resnet on 3 domains and alternate the target domain to test which domains are more difficult when it comes to generalisation and which model is able to perform better. For Shape Bias we will be measuring the ratio of correctly identified images in conflicted dataset with correctly classified images in normal dataset. All other tests will use normal accuracy metric for analysis. We will also try to visualise the feature space just before classification head using t-SNE to map our feature space to 2D vectors.

2.2. Results

After fine tuning both models for 2-3 epochs the base accuracy for ViT-S/16 was approximately 97% while for ResNet it was approximately 96%. We ensured that the baseline performance is comparable to make the subsequent tests more fair.

2.2.1. Semantic Biases

For color bias we generated a gray scale dataset from STL-10 test set and tested both model's performance. The ViT's accuracy dropped to 92% while the ResNet's accuracy dropped to 91%. Since ViT is generally more biased towards shapes while CNN is more color biased the ViT performed slightly better. While we were expecting a significant performance difference for both models, since the dataset was very small both models had learned the nuances of data and performed good.

For shape bias we generated a cue conflicted dataset from STL-10 test set and analysed performance. To analyse shape bias we use shape bias % as explained above and the ViT had shape bias % of 21 while ResNet reached 38%. This was unexpected since the ViT is considered more shape bias than CNN. Since the dataset was limited our model had learned shortcuts and performed poorly when a change occured. We examined the performance closely and found that ViT had a decent performance on classifying images of cars and cats which have a more defined shape than others. Upon examining the conflicted dataset we realised that a few images had lost some of its shape leading to worse performance since ViT depends on shapes. We then tried a more complex model, ViT-B/16 from google to check whether training and limited dataset was the problem. Upon training it for the same number of epochs we reached a shape bias % of 27. This shows that ViT requires more training to learn the global contexts when compared to ResNet50. ResNet50 requires less training and had more parameters as compared to ViT-S/16 which could be a potential reason for better performance but overall training and a larger dataset would allow a better comparison between ViTs and CNNs.

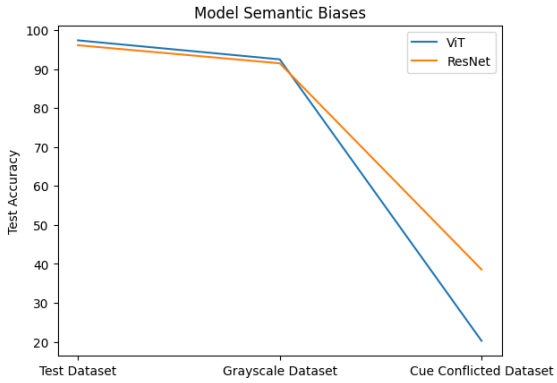

Figure 1. Accuracies to compare semantic biases in discriminative models.

2.2.2. Locality Bias

We first analysed translational invariance on both models. Shifting images had little to no difference in performance of both models which is because both models had learned the patterns in data because training data had only 500 images per class label.

We then masked out patches in the images and ViT got an accuracy of 85% while ResNet dropped to 45%. This is because CNNs tend to look for edges and local textures which got lost in masked dataset. ViT was able to understand the global context of the image and performed significantly better.

Our final test was with a permuted set of images. Since ViTs rely on global performance while CNNs can still fire on local matches our CNN had slightly better

performance when we shuffled our image using patch size of 16x16. However, as we increased patch size to 28x28 the ViT was able to pick the global context and started to perform much better. The overall performance indicates that CNN based models are more locality biased while ViTs focus on global context.

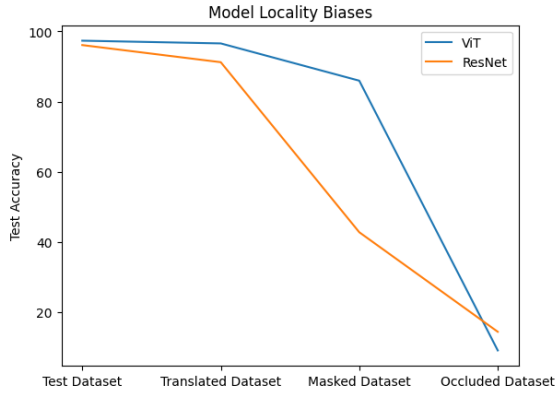

Figure 2. Accuracies to compare locality biases in discriminative models.

2.2.3. Domain Generalisation

For domain generalisation we picked 3 source domains and 1 target domain and the trends for both models were very similar. They achieved least test accuracy when sketch was the target domain. This is because sketch is most different in terms of semantics when compared to other datasets, therefore, domain generalisation for it was more difficult. In this domain CNN was able to achieve 50% accuracy while ViT was able to reach only 20%. This could be due to lack of training since ViT require intense training to understand shapes. Moreover, the ability of CNNs to pick edges better than ViT was another reason for better performance, since, sketch domain has no other information. On other test domains ViT was able to give a better performance [3], showing the ability of ViTs to focus more on shapes than textures/colors of images when compared to CNNs.

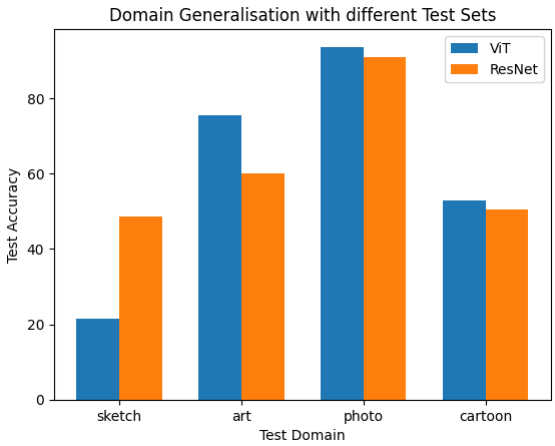

Figure 3. Domain generalisation on different test sets

2.3. Analysis

In this section we analysed inductive biases in discriminative models and how they affect generalisation. We identified the ViT's ability to understand global context by its performance on the masked dataset where CNNs lost much of their ability to distinguish. This aligns with existing findings, but some of our other results, particularly for semantic biases, were not as consistent with prior work. A major reason for this is the limited dataset size and the short training time of only 2–3 epochs, which likely led both models to exploit shortcuts rather than fully learn global or shape-based cues.

For locality biases, the results were in line with expectations: ResNet relied heavily on local features and struggled with masked inputs, while ViT was more robust. However, the magnitude of the gap (85% vs 45%) further reinforce the idea of inadequate performance due to the small scale of STL-10 rather than purely architectural differences. Similarly, in the shuffled dataset experiments, patch size played a large role, highlighting how design choices in ViTs can strongly influence their behaviour, using smaller patch sizes made local features dominant, making CNNs perform better, while, increasing patch size made ViT outperform since it was able to pick global cues.

In the case of semantic biases, our conflicted dataset generation process removed key shape information from some of the images, penalising ViT disproportionately which explains why ResNet appeared more shape-biased in our results. These results, apparently, contradicts the findings of (Robert Geirhos, 2022) where CNNs proved to be more biased towards texture instead of shape. However, when we take a deeper dive we can observe that our findings suggests model overfitting due to small size of STL-10 dataset. Moreover, CNNs require less compute to achieve adequate performance while ViTs require more compute and data to understand the global context properly for which a few

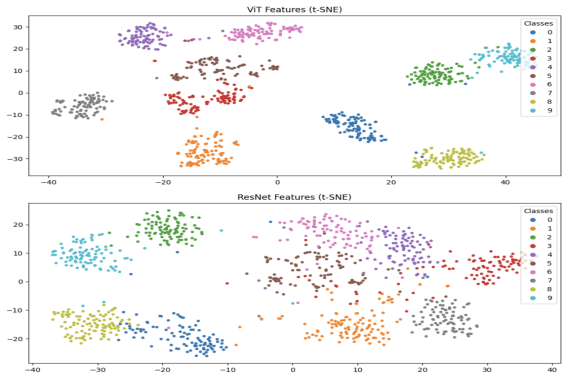

Figure 4. ViT vs ResNet Feature Space

epochs were not enough to train it. This allowed CNN to perform better than ViT. Both models reached a baseline accuracy more than 95% on only 2-3 training epochs which clearly indicate that models had learned shortcuts since train and test sets have only 500 and 800 per class images respectively in STL-10 dataset. To reproduce results, we would require more data and compute for example using Stylized-ImageNet against normal ImageNet would give more reliable comparisons.

When analysing domain generalisation, both models performed poorly on the sketch domain, with CNNs outperforming ViTs due to their ability to recognise edges in images. On other domains ViTs were able to perform better, but the overall accuracy remained low. Sketch domain is reletively difficult to use as test set due to vast differences between sketch and other domains but performance on other test domains reinforce that ViTs can perform better at shapes when compared to ResNet.

Overall, our experiments suggest that while architectural inductive biases are important, the training methods, dataset design, and evaluation choices play a critical role in the performance. Both models achieved high baseline accuracy on normal test data, but under stress tests their performance dropped significantly. This can be seen from the tight clustering in feature visualisations [4] which indicates slight over fitting of models on test dataset. The inconsistencies we observed point less to contradictions in the literature and more to insufficient training and limited data. With longer training, larger and more robust datasets, and standardised evaluation metrics, we expect the trends for shape and texture bias to align more closely with prior findings.

3. Generative Models (VAE vs GAN)

3.1. Methodology

In this work, we train a convolutional VAE and a DCGAN on the CIFAR-10 dataset for a total of 50 epochs each, which serves as the benchmark for all the following experiments we will conduct. Both models are made up of five convolutional blocks each, with a latent dimension of 128.

For optimization, we use the Adam optimizer with learning rates set to 1×10^{-3} for the VAE and 2×10^{-4} with $\beta = (0.5, 0.999)$ for the DCGAN. A step learning rate scheduler (`StepLR` with step_size=10, $\gamma = 0.5$) is used in both. The GAN is trained with `BCEWithLogitsLoss`, while the VAE is trained with a weighted sum of mean squared error (MSE) and KL divergence, i.e.,

$$\mathcal{L}_{\text{VAE}} = \alpha \text{MSE} + \beta \, \text{KL}.$$

Model evaluation is conducted along multiple axes: sample fidelity and diversity are assessed using Fréchet Inception Distance (FID) and Inception Score. The latent structures of the models are analyzed through t-SNE visualizations. Overall quality of the generated images are also checked visually and by putting it through varying experiments like interpolating it from one latent dimension to another. Out-of-distribution (OOD) robustness is also evaluated using VAE reconstruction errors and GAN latent stress tests. To ensure reproducibility, model checkpoints and generated sample grids are saved every 5 epochs.

3.2. Results

3.2.1. Vision Quality and Diversity Check

Training the GAN occasionally showed instability, with the discriminator overpowering the generator in early epochs before balancing out (yet still continuing to have spikes of high loss here and there). The VAE's training however was more stable but converged to blurry reconstructions. This contrast reflects the well-known trade-off: GANs can be unstable but yield high-fidelity samples, whereas VAEs are stable but produce blur. However, after 50 epochs of training, we observed the following qualitative outputs.

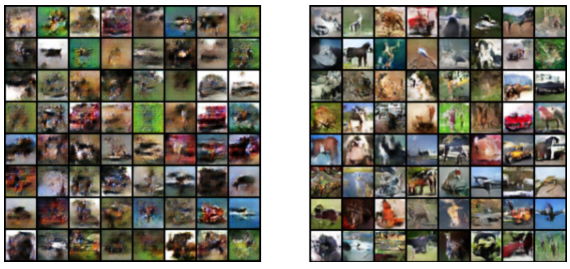

Figure 5. Initial and Final GAN Visualization

For the GAN, early samples were very blurry and lacked recognizable features. By the end of training, the outputs contained clearer features corresponding to different CIFAR-10 classes. We also observed that the GAN continued to

produce varied samples across classes, with no visible signs of mode collapse.

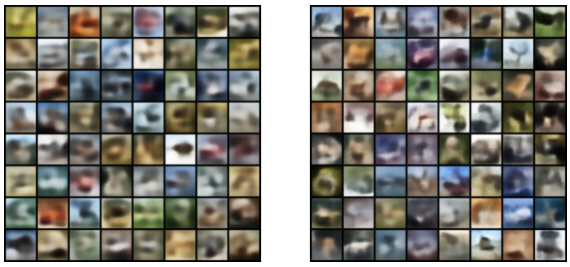

Figure 6. Initial and Final VAE Visualization

For the VAE, outputs remained blurry throughout training. While outlines of the original images were sometimes visible, finer class-specific details were not recognisable and the images were overall quite blurry. Side-by-side comparisons with the original images were needed to identify the intended object in many cases.

3.2.2. Metrics

We further quantified the performance of both models using standard metrics. The results are summarized in Table 1.

Model	Reconstruction Error (MSE)	Inception Score (IS)	FID
VAE	0.0486	–	405.7
GAN	–	4.2977	422.0

Table 1. Metrics for VAE and GAN after 50 epochs of training.

3.2.3. Interpolation

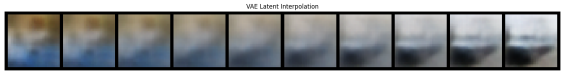

Figure 7. Interpolation of VAE

Figure 8. Interpolation of GAN

For the VAE, interpolations between latent codes produced smooth intermediate images, with gradual transitions between endpoints (Figure 7). For the GAN, interpolations showed sharp images at the endpoints but had a very sharp change between them (Figure 8). It was not at all very smooth.

3.2.4. Latent Feature Visualization

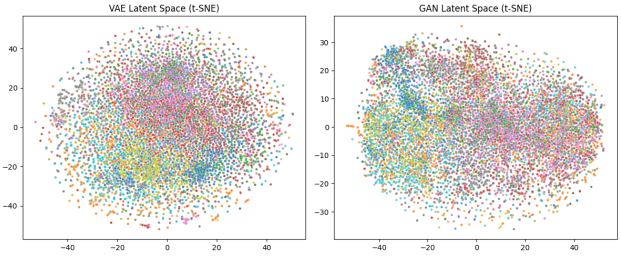

Figure 9. t-SNE projections of VAE latent embeddings and GAN latent samples.

t-SNE projection of the VAE's latent embeddings showed weak clustering patterns, with some visually similar classes (e.g., cars and trucks) appearing closer together. The same projection on GAN latent samples showed only faint grouping tendencies and lacked clear separation between classes.

3.2.5. Out-of-Distribution Behaviour

Figure 10. In-distribution Reconstructions of VAE

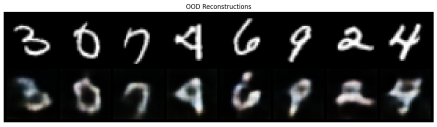

Figure 11. Out-of-distribution Reconstructions of VAE

When evaluated on MNIST digits as OOD inputs, the VAE produced reconstructions that resembled distorted CIFAR-10 images with added color and blurred edges (Figure 11).

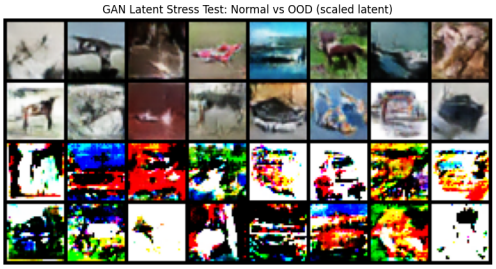

Figure 12. GAN generations under ID vs. OOD input scaling

For the GAN, scaling the latent vector to mimic OOD inputs resulted in sharp but oversaturated images from the CIFAR-10 distribution, unrelated to the MNIST digit input (Figure 12).

3.3. Analysis

3.3.1. Visual Quality and Diversity

After 50 epochs of training, we observed distinct differences between the VAE and GAN in terms of visual quality and diversity. The GAN initially produced very blurry outputs, where it was difficult to discern the intended objects. However, as training progressed, the generated images became increasingly coherent, with recognizable features forming across different classes. For example, the horse image in the [2,2] position of Figure 5 initially lacked limbs but gradually developed horse-like features until it became clearly identifiable. Importantly, the GAN did not exhibit signs of mode collapse and consistently generated diverse samples across the dataset. Despite architectural constraints, the final outputs demonstrated reasonably good quality across all classes.

In comparison, the VAE results were less promising. While it produced rough outlines that loosely resembled the original inputs, the generated samples remained blurry throughout training. Recognizable structure was often visible only when comparing reconstructions side-by-side with the originals, underscoring the well-known limitation of VAEs in capturing fine detail.

3.3.2. Quantitative Metrics

To further quantify these observations, we evaluated both models using standard metrics. For the VAE, the average reconstruction error was 0.0486 MSE per pixel, consistent with its ability to capture overall structure but not finer details. For the GAN, which does not support direct reconstruction, we instead measured the Inception Score (IS) on 1000 generated samples, achieving 4.2977, reflecting a balance between diversity and class consistency. Finally, when comparing both models using the Fréchet Inception Distance (FID), the VAE achieved a slightly lower score (405.7) than the GAN (422.0). This is not at all what literature (Bond-Taylor et al., 2021) suggests (that being that GANs generally produce sharper and thus lower-FID outputs), and highlights the limitations of our simple GAN architecture.

3.3.3. Latent Space Interpolation

When examining latent space continuity through interpolation, the VAE displayed smooth transitions across the latent manifold, with each intermediate sample appearing plausible (Figure 7). This indicates that the VAE successfully learned a continuous, structured latent space where linear interpolations map to semantically meaningful transformations. The GAN interpolations (Figure 8), while capable of producing sharp outputs at the endpoints, often included abrupt or incoherent intermediate images. This reflects the absence of explicit regularization in GAN latent spaces, which makes them less consistent and interpretable compared to VAEs.

3.3.4. Latent Feature Representation

The latent representations of both models also reflected this difference. For the VAE, t-SNE projections of latent embeddings revealed weak but noticeable structure, with semantically similar classes (e.g., cars and trucks) appearing closer together (Figure 9). This behavior aligns with the VAE's encoder, which explicitly maps inputs to a structured latent space. The GAN's latent codes, sampled directly from the prior, showed only faint clustering tendencies. While this suggests the generator implicitly imposes some organization, the structure was far less interpretable than in the VAE. This reinforces the difference between VAEs' explicitly regularized latent variables and GANs' emergent, loosely structured representations. Overall, both of their projections were quite weak and did were not reflective of the true probability distributions of the dataset and either more training or a stronger model (or both) is needed.

3.3.5. Out-of-Distribution Behavior

We also examined the behavior of both models on out-of-distribution (OOD) data, using MNIST digits as input. For the VAE, reconstructions revealed its inductive bias: grayscale digits were "reinterpreted" as CIFAR-like images, often with added color and blurred edges (Figures 10, 11). This shows how the VAE assumes every input belongs to the training distribution, even when it does not. The GAN, on the other hand, cannot reconstruct inputs. Instead, when fed scaled latent codes to mimic OOD inputs, it produced sharp but oversaturated CIFAR-like samples (Figure 12). These outputs bore no relation to the original digits, underscoring the GAN's limitation in tasks like anomaly detection.

3.3.6. Overall Comparison

Overall, these findings reflect the expected trade-offs between VAEs and GANs. The GAN generated sharper and more coherent images but with less latent structure and weaker generalization outside the training distribution. The VAE, in contrast, produced blurrier samples but demonstrated stronger latent continuity and interpretability, as well as the ability to reconstruct inputs. These results align with the general understanding that GANs prioritize fidelity, while VAEs emphasize diversity and structured representation.

4. Contrastive Models (CLIP)

4.1. Methodology

We chose OpenAI's CLIP (ViT-B/32) (Radford et al., 2021) to investigate the extent of shape and texture biases and the implications of unique multimodal and contrastive biases when using image-text based contrastive models for image classification.

Our ResNet50 model provided a baseline with which to compare CLIP's performance on equivalent or analogous tasks. For most of the analysis, we again used the STL-10 dataset.

The zero-shot pipeline for image classification involves generating text prompt embeddings according to the dataset classes and image embeddings for the sample to make predictions for. The final prediction is the class with the highest cosine-similarity text feature to the image features. Where possible, we directly used or adapted prompts from the original CLIP paper (Radford et al., 2021). Examples of these include "a photo of the class.", "a blurry photo of a {class}.", and "a bad photo of the {class}.". This prompt ensembling (versus single prompts such as "a {class}") allowed us to get meaningful and more generally applicable text embeddings. Images were processed by the CLIP preprocessor and features were extracted using the CLIP image encoder as-is.

CLIP's performance on domain-shifted classification was analyzed using zero-shot performance on PACS sketch, art, and cartoon domains. This follows standard practice when evaluating domain shift performance of other models, including our baseline, which were trained only on photographs. However, CLIP is unique because its training set, although primarily photographs from the internet, also included images from other domains. Further, we evaluated the effectiveness of specifying the target domain in the text prompts, generating results for both this scenario as well as simple "a {class}" prompts. Both of these are compared to our baseline ResNet's performance in two scenarios: trained on PACS photo only, and trained on every PACS class except one test class.

As well as image-to-text (I2T), CLIP's multimodal architecture allows for text-to-image (T2I) retrival. This capability was tested by querying a small (20 image) subset of PACS containing samples from different domains and classes.

CLIP's clustering of image features across domains and classes was assessed using tSNE dimensionality reduction to visualize the feature space. For this, a mixed dataset containing CIFAR-10 and PACS samples was used. This was compared with the baseline ResNet features extracted at three different depths of the network.

The severity of shape and texture bias in CLIP was determined using zero-shot performance on the same cue-conflicted version of STL-10 tested on our ResNet50 model. This allowed us to judge and compare CLIP's specifically on what semantics the model learned during its training as well as its ability to generalize across stylized domains. Here, we again measure shape bias as the ratio of correctly identified images in the cue-conflicted dataset to the correctly identified images in the base, non-cue-conflicted dataset.

Lastly, the robustness of CLIP to significant perturbations (namely, stylistic renditions, noise, and blur) on singular samples was compared to that of our baseline model. This allowed us to gauge the stability and capacity of CLIP, which can have implications in real-world use cases i.e. object detection in noisy, blurry CCTV images.

4.2. Results

4.2.1. Zero-Shot Performance

ResNet50	CLIP (simple prompts)	CLIP (prompt ensembling)
96.10%	96.25%	96.78%

Table 2. Accuracy on STL-10 test set of ResNet50 (trained on STL-10) and CLIP (zero-shot).

CLIP exhibits highly impressive zero-shot accuracy on STL-10 data, as shown in Table 2. Without any modifications to model architecture or parameter-tuning, CLIP matches the performance of our baseline ResNet50 model which was trained on the STL-10 test set. This dataset is entirely unseen to the CLIP model, indicating that CLIP generalizes to OOD data with ease.

Even with simple prompts, i.e. the prompts being just the class names, CLIP demonstrated accuracy beyond that of our specialized ResNet50. Prompt ensembling using the CLIP paper STL-10 prompts (i.e. "a photo of a .", "a photo of the .") further led to a 0.53% increase in accuracy – an effect which is difficult to ignore when accuracy is already in the higher 90s.

The per-class difference in accuracies allows deeper analysis of CLIP performance and its differences to ResNet. As shown in Table 3, both models achieve similar accuracies on most of the STL-10 classes. However, CLIP is able to improve greatly where ResNet suffers – this is highlighted especially by the *dog* class, where CLIP achieves an accuracy increase of 18.25%. This overcoming of particular challenges showcases the benefits of CLIP over CNN models. The increase in accuracy on *dog* coincides also with large increases on *deer* and *horse*. One explanation of this is that the ResNet could have been mispredicting these classes amongst each other, and CLIP has not faced that same difficulty.

However, it is noteworthy that CLIP is not always more

Class	ResNet50	CLIP	Difference
airplane	94.75%	99.12%	+4.37%
bird	95.75%	99.75%	+4.00%
car	97.88%	95.75%	−2.13%
cat	88.25%	84.25%	−4.00%
deer	90.88%	98.12%	+7.24%
dog	77.88%	96.12%	+18.24%
horse	91.00%	98.62%	+7.62%
monkey	98.88%	97.25%	−1.63%
ship	99.25%	99.88%	+0.63%
truck	96.00%	98.88%	+2.88%

Table 3. Per-class accuracy on STL-10 test set of ResNet50 (trained on STL-10) and CLIP (zero-shot).

accurate and should not be considered to be without flaws as, in 3 of the 10 classes, the baseline model was up to 4% more accurate. There may be deeper model intricacies which lead to these specific deficiencies in CLIP, and this phenomenon is well-documented in literature (Shao et al., 2023).

4.2.2. Domain Generalization

Test Domain	ResNet 1	ResNet 2	CLIP (simple)	CLIP (domain)
PACS Art/Painting	60.89%	63.77%	95.02%	94.87%
PACS Cartoon	27.30%	55.08%	97.53%	97.53%
PACS Sketch	33.80%	52.28%	83.69%	86.56%

Table 4. Accuracy on PACS non-photo domains. ResNet 1: trained on PACS photo. ResNet 2: trained on all PACS except test domain. CLIP simple prompt: "a {class}". CLIP Domain specified: "a {domain} of a {class}" with ensembling.

Our results show that CLIP achieved highly on all chosen domain-shifted sets. Table 4 shows CLIP accuracy on PACS Art/Painting and PACS Cartoon is very impressive – comparable to that on STL-10. Accuracy on PACS Sketch is $11 - 13\%$ lower than this, indicating relative struggle. Including the domain in the prompt for CLIP classification does not affect performance on Art/Painting and Cartoon, but it is able to recover 2.87% of the lost accuracy on Sketch, which is notable. This shows that prompt ensembling can be helpful in cases where CLIP does not perform well on domain-shifted data.

The two ResNet models, albeit fine-tuned on different portions of the PACS dataset, do not perform nearly as well as CLIP. The highest accuracy achieved here was just 63.77% for a model trained on PACS Photo, Cartoon, Sketch evaluated on PACS Art/Painting. Alongside the rest of the results, this shows that the ResNet is unable to generalize easily to domain-shifted data, even after being exposed to renditions of the same classes.

Figure 13. CLIP image retrieval from shown query prompts

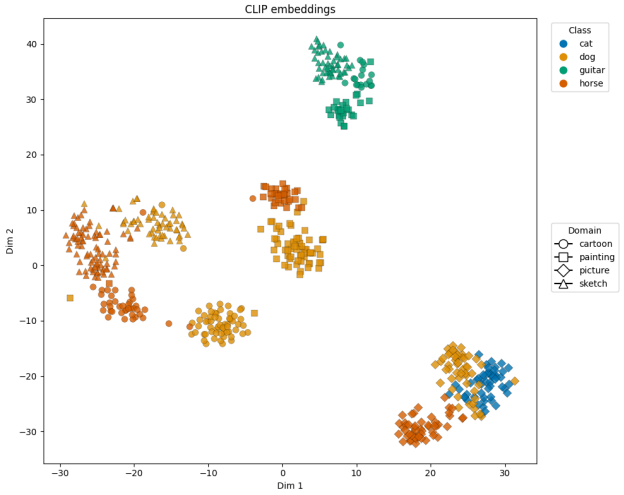

Figure 14. tSNE of CLIP image embeddings

4.2.3. Image-Text Retrieval

When tasked with retrieving images corresponding to query prompts (contain domain and class information), CLIP was able to do so correctly for most samples, highlighting that these domain-specifying queries allowed CLIP to tailor its text embeddings and maximize similarity with a particular image. The results in Figure 13 showcase CLIP's rich joint vision-language representation as the model was able to pick out a painting of horses from a collection of horses of other domains, all after being prompted once. Still, it is imperfect as seen in the sketch of the dog, which was misunderstood as a cartoon by CLIP, indicating that differentiating between two non-photo domains may be a shortcoming of CLIP.

4.2.4. Image Representation Quality

Using our mixed CIFAR/PACS dataset, we found that CLIP clusters image features (Figure 14)in two ways. First, there is broader grouping based on domain (primarily bottom right for photos), which differs from claims that CLIP is always able to group class samples together regardless of domain and points towards existence of domain gaps. Then, within these groups, there is more specific, class-based clustering, indicating purposeful bridging of class gaps. This separability of features is the root of CLIP's high accuracy.

Our plot includes two particularly distinct classes: *guitar*

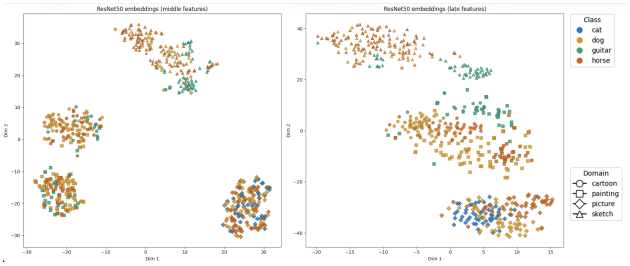

Figure 15. tSNE of ResNet50 feature spaces. Left: layer 3. Right: Final avgpool.

Figure 16. ResNet and CLIP predictions on perturbated PACS images. True labels: bird, monkey, truck (L-R)

(PACS non-photo only) and *cat* (CIFAR only). These further prove our analysis - *cat* is grouped with other photos and *guitar* is far away, closer to other non-photos. The distance between *guitars* (regardless of domain) and the other PACS samples can be attributed to CLIP's ability to recognize and cluster based on shapes, as guitars have unique shapes.

With ResNet (Figure 15), layer 3 features show strong domain-based clustering – in fact, they are similar to the CLIP features in that the distances between images and all other domains are quite high. This shows that both ResNet and CLIP feature domain gaps. In the later ResNet features, we see the model attempt to form linearly seperable clusters based on class. However, it is not able to do as well as CLIP, which is why we see lower accuracy here on domain-shifted datasets. The key difference in interpreting the results is CLIP's use of text features. While CLIP also features domain gap, we can conclude (due to high accuracies in classification), that these gaps are also found in the text embeddings of prompts containing the domain information. This can be one way of effectively "bridging the domain gap" in CLIP.

4.2.5. Inductive Biases

ResNet50	CLIP (simple prompts)
38.55%	77.61%

Table 5. Shape bias when using cue-conflicted STL-10 test set.

On a stylized/cue-conflicted dataset, CLIP achieved a shape bias higher than our ResNet50 model by a factor of 2 (Table 5). This is explained by CLIP's use of attention layers (from ViTs) in encoders, which rely heavily on shapes in images, and less so on texture and other semantics. This high accuracy on stylized images shows that CLIP is resilient to changes in presentation of data, and this may be attributed to it's large training data size.

CLIP not only exceeded ResNet, it even had a much higher shape bias than our ViT. Ignoring dataset constraints, this could suggest that the multimodal training of CLIP has further enhanced its semantic biases (in this case, shape) beyond that of ViTs.

4.2.6. Robustness and Limitations

In our testing, CLIP was also able to prove its robustness when being tested on difficult data featuring perturbations such as blur, noise, and filters. CLIP was able to predict the correct class or a meaningfully related class for many of these samples (some of which even humans may struggle with), as shown in Figure 16. This shows that CLIP image embeddings continue to be meaningful and similar to their positive text counterparts despite alterations to the image, providing basis for real-world use where data is seldom as expected. In our testing, one of the few images CLIP miscategorized was the noisy truck in Figure 16. Even still, CLIP demonstrated robustness by predicting *car*, which is a reasonable prediction.

In contrast, our ResNet50 struggled to predict most of these samples correctly, and in certain cases, the predictions were not meaningfully related to the true label. Again, using the noisy truck as an example, the ResNet model's top prediction was cat, an unrelated class. Notably, both ResNet and CLIP classified images stylized as cartoons at about the same rate as they would regular STL images. This seems quite intuitive, as many of the semantic features are preserved in this setting, whereas with blur or noise, these may become difficult to discern.

4.3. Discussion

CLIP's multimodal architecture has translated into crystal-clear improvements in generalization across domains in comparison to other state-of-the-art machine learning models. The combination of text-image pairing and training on such a large variety of data (allowing for one-shot with high accuracy) results in a model which is highly accurate and applicable for many use cases.

High shape bias is a highly desirable characteristic and CLIP does an excellent job of mimicking human-like shape recognition capabilities compared to other vision models.

Although we can conclude that the higher shape bias is beneficial, our investigation did not find that this came at the cost of ignoring texture, color, background, and other semantics (which would not be desirable). A more focused study may help in completing this picture.

We know that CLIP generalizes at a level not easily achieved with ViTs even though they share the same underlying architecture. It must be CLIP's training data volume and variety which has injected in it the type of biases that the model shows. We encourage this not to be understood as conventional fine-tuning because the model does not become limited to a narrower domain. Instead, learning from this breadth of data allows CLIP to aready be "fine-tuned" for almost any domain.

It is worth acknowledging that not all of CLIP's biases are desirable. Although not in our scope, studies have shown that training from internet images has resulted in gender, racial, and age bias in CLIP. These present a set of unique challenges which literature is continuing to address (Zhang et al., 2024).

5. Conclusion

In summary, our investigations find that semantic and inductive biases differ across discriminative, generative, and contrastive models by severity and causes. Each model's training and architecture context leads to assumptions which we have been able to show in a variety of scenarios. These assumptions prove useful in some contexts and less so in others. No one model has proven it's fitness or adaptability to every task to be better than all other models. Instead of trying to achieve a "one-for-all", it may prove more fruitful to continue to methodically incorporate appropriate biases in design and training in order to maximize generalization across model families while retaining in-domain capabilities.

References

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models, 2021. URL https://arxiv.org/abs/2103.04922.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, Virtual, 2021. PMLR. URL https://proceedings.mlr.press/v139/radford21a.html.

Robert Geirhos, Patricia Rubisch, C. M. M. B. F. A. W. W. B. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. 2022.

Shao, J.-J., Shi, J.-X., Yang, X.-W., Guo, L.-Z., and Li, Y.-F. Investigating the limitation of CLIP models: The worst-performing categories, 2023. URL https://arxiv.org/abs/2310.03324.

Zhang, H., Guo, Y., and Kankanhalli, M. S. Joint vision-language social bias removal for clip. *arXiv preprint*, November 2024. URL https://arxiv.org/abs/2411.12785. arXiv:2411.12785 [cs.CV].