Advanced ML Assignment 1

Muhammad Huraira Anwar Adeen Ali Khan

1. Abstract

State-of-the-art discriminative and contrastive models perform well when trained and tested on similar data distributions but struggle significantly with out-of-domain (OOD) data. In this work, we investigate the domain generalization capabilities and inductive biases of three models: Vision Transformer (ViT), Contrastive Language-Image Pretraining (CLIP), and Visual Geometry Group (VGG-16). We explore the models' ability to generalize to OOD data using datasets like PACS and SVHN, which exhibit covariate and semantic shifts, and further evaluate their inductive biases—specifically their reliance on shape, texture, and color—by creating custom datasets from popular sources like CIFAR-10 and COCO, introducing controlled variations in these visual cues. We also evaluated the models' reliance on local versus global information by introducing localized noise, global style changes, and scrambling image patches, gaining insights into their locality biases. Additionally, we assessed the efficiency and accuracy of diffusion models as zero-shot classifiers, using their denoising ability to predict label likelihoods. Under our GPU and time constraints, diffusion models performed poorly compared to CLIP in zero-shot classification tasks, demonstrating weak generalization to unseen domains. Finally, we constructed six different single-layer models using different combinations of convolution and attention to further study how they capture local and global information, linking these insights to the models' locality versus global bias through visualization of feature and attention maps.

2. Introduction

Deep learning models have achieved remarkable success in various computer vision tasks, particularly when trained and tested on data sampled from similar distributions. State-of-the-art discriminative and contrastive models, such as attention-based models like ViTs, CLIP, and convolutional architectures like VGG16, perform exceptionally well in these in-domain settings. However, they face significant challenges when exposed to out-of-domain (OOD) data, which is critical for real-world deployment where models must generalize beyond their training distributions.

Domain generalization is a crucial aspect of enhancing model robustness to such OOD scenarios. In this work, we investigate the domain generalization capabilities and inductive biases of three popular models—ViT, CLIP, and VGG16. Specifically, we evaluate how well these models handle domain shifts, such as covariate and semantic shifts, by leveraging datasets like PACS and SVHN. PACS introduces shifts in style and representation (e.g., Photos, Art paintings, Cartoons, Sketches), reflecting a covariate shift, while SVHN introduces semantic shifts that pose unique challenges for model generalization.

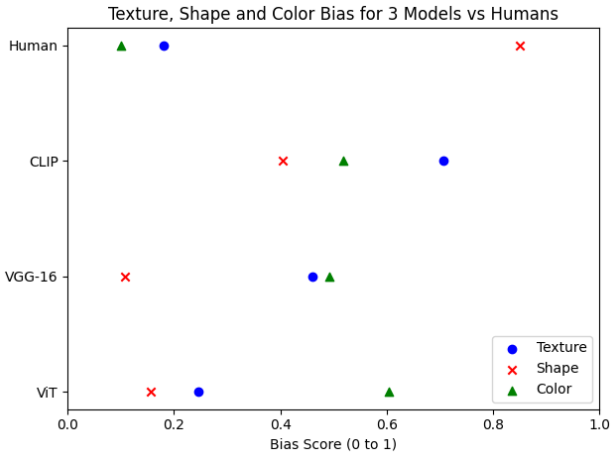

Figure 1. An evaluation of all 3 models with comparison to humans on how they perceive textures and shapes and how their inductive biases, which are used to classify objects, compare with each other. It strongly highlights the shape bias of humans, the preference of local dependencies by CNNs like VGG-16 and a greater understanding of global structures by attention-based models like ViTs, and a balanced neutrality of contrastive models like CLIP which value both local and global information almost equally.

To better understand the inductive biases of these models, we further examine their reliance on shape, texture, and color information. We created custom datasets derived from popular sources like CIFAR-10 and COCO, modifying the images to emphasize different visual cues—such as isolating shape, introducing texture alterations, and changing color distributions. This approach allows us to systematically analyze whether a model is more biased towards certain features, which can significantly affect its robustness to distributional changes.

We also evaluated the models' reliance on local versus

global information by introducing localized noise, applying global style changes, and scrambling image patches. This analysis provided insights into their locality biases and how well they handle different types of alterations affecting either small regions or the entire image structure.

In addition to evaluating discriminative and contrastive models, we assessed the capabilities of diffusion models for zero-shot classification. By leveraging their denoising mechanism to predict label likelihoods, we aimed to determine their effectiveness in handling OOD data. However, under our GPU and time constraints, diffusion models performed poorly compared to CLIP, highlighting their weak generalization capabilities in zero-shot classification tasks.

Lastly, we aimed to gain insights into how models capture local and global information by constructing six single-layer models that combine convolution and attention mechanisms. By visualizing feature and attention maps, we analyzed the extent to which these models focus on different aspects of the input, providing a better understanding of their internal representations and how they process visual data.

This study not only highlights the limitations of current state-of-the-art models in handling OOD scenarios but also provides a comprehensive evaluation of their inductive biases, contributing towards developing more robust models for real-world applications.

3. Methodology

Task 1: Zero-Shot Classification using Diffusion Models

For zero-shot classification, we utilized diffusion models by conditioning the denoising process on specific class labels to generate class-specific representations. We began by applying noise at various randomly sampled timesteps to a latent image that represented a particular class. This noisy latent was then gradually denoised to reconstruct the original latent image. During the denoising process, at each timestep, we computed the L2 norm difference between the original latent representation and the denoised version.

To make the classification more accurate, we applied a weighting scheme to these L2 norm differences. Specifically, the differences were weighted based on the timestep, with higher timesteps receiving lower weights. This was because, at higher timesteps, the image becomes closer to its original form, and therefore requires less noise removal compared to earlier stages. We then took the mean of these weighted scores across all timesteps for each class and stored it as a measure of how well the model could denoise the latent image for that class.

This process was repeated for all possible classes, and at the end, the class with the lowest mean weighted score was selected as the predicted class using the argmin operation. By doing so, we identified the class for which the denoising process was most effective, indicating the closest match for the given image representation in a zero-shot setting.

Task 2 & 3: Fine-Tuning and Testing on IID and OOD Data

We fine-tuned the models on the CIFAR, PACS, and SVHN datasets to evaluate their performance on both IID (in-distribution) and OOD (out-of-distribution) data. This setup allowed us to analyze how well the models generalized to new, unseen data and environments.

For the PACS dataset, we used a domain generalization approach, where the model was trained on one domain (e.g., Photos, Art, Cartoons, or Sketches) and then tested on all the other domains. This approach enabled us to measure how effectively the model handled domain and covariate shifts, as each domain presents a different style or representation, differing significantly from the training domain.

For the SVHN dataset, our focus was on evaluating the model's robustness to semantic shifts. The SVHN dataset features images of house numbers, which have semantic distributions that differ from datasets like CIFAR. By training the model on one distribution and testing it on another, we assessed how well it could generalize to shifts in semantic content, giving insight into its ability to adapt to new, conceptually different data types.

This combined evaluation on both domain generalization and semantic robustness provided a comprehensive view of each model's capability to generalize effectively to diverse OOD conditions.

Task 4 & 5: Evaluating Inductive Biases

For Tasks 4 5, we conducted an in-depth evaluation of the inductive biases of the models by testing them under controlled shifts, aiming to better understand their inherent strengths and limitations when dealing with various types of alterations in the data. To facilitate this evaluation, we created custom datasets specifically designed to emphasize different visual features, such as texture, color, shape, scrambled, and noisy images.

To evaluate semantic biases, we introduced consistent color shifts throughout the dataset, allowing us to examine how well the models could adapt to changes in color information—a critical factor in understanding high-level semantics. Specifically, we created a color-biased dataset where distinct color transformations were applied to the images, effectively testing the reliance of each model on color as a distinguishing feature. This allowed us to determine the extent to which the models relied on color information to make their predictions, and whether such changes affected their classification performance.

Additionally, we created a texture-biased dataset by apply-

ing style transfer to all images using a pre-trained VGG19 model, incorporating the "Starry Night" style. This transformation aimed to evaluate the models' performance when exposed to significant changes in visual style, testing their ability to handle altered texture and visual features while still recognizing the underlying content. Furthermore, we employed the SAM (Segment Anything Model) to generate segmented images and extract silhouettes from the COCO dataset, effectively creating a shape-biased dataset. This dataset allowed us to test the models' ability to recognize objects based on their shape alone, devoid of color or texture cues, providing insights into how well each model could handle high-level content shifts with minimal reliance on detailed features.

To evaluate locality biases, we injected localized perturbations by adding Gaussian noise to the center of each image, effectively creating a noisy dataset. This approach helped us determine the models' sensitivity to specific localized disruptions and their ability to maintain accuracy despite such noise in certain parts of the image. Additionally, we scrambled the images by dividing them into smaller sections and randomly rearranging those sections, creating a scrambled dataset. This process was intended to simulate a disruption in the global view of the image, testing the models' capability to understand the overall context of the content when the spatial relationships between parts of the image were lost.

By conducting these controlled tests using custom datasets focusing on texture, color, shape, scrambled, and noisy images, we systematically identified and compared the inductive biases of the models. The color-biased dataset showed the models' sensitivity to changes in color information, while the texture-biased dataset highlighted reliance on local patterns. The shape-biased dataset assessed the models' ability to capture global structure, whereas the noisy and scrambled datasets tested their handling of local noise and disruptions to the global context.

These experiments provided deeper insights into each model's generalization capabilities and robustness in the face of various visual transformations. The findings were crucial in identifying both strengths and limitations, offering guidance on improving overall generalization and OOD performance, and adapting models to handle a wide range of real-world visual variations effectively.

Figure 2. An illustration of how we are using an image from the COCO dataset and applying SAM model on it to retrieve it's silhouette which is being used later for shape bias evaluations.

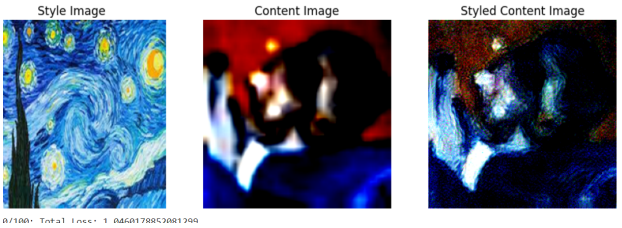

Figure 3. An illustration of how we are using an image from the CIFAR 10 dataset and applying VGG-19 style transfer model on it with 'starry night' image to impose its texture on the input image. This is being used later for texture bias evaluations.

Task 6: Evaluating Single-Layer Models

In Task 6, we evaluated single-layer models by implementing and visualizing different mathematical operations that combined convolution and self-attention. These operations helped us compare how various combinations of local and global information influenced feature extraction, highlighting the strengths and weaknesses of each approach.

We analyzed six types of operations that used different combinations of depthwise convolution and self-attention. Convolution-based operations focused on capturing local patterns within a defined receptive field, whereas self-attention expanded the receptive field globally, enabling the model to understand broader dependencies. By combining these mechanisms in different ways, such as post-normalization and pre-normalization, we explored how information was aggregated, affecting how receptive fields were formed.

To further explore these effects, we implemented a single-layer model for each operation and visualized the feature maps using a sample image. Models utilizing convolution components effectively captured localized details, while those incorporating self-attention captured global relationships. The visualizations demonstrated how each model emphasized different aspects of the image, revealing the inherent limitations and strengths of single-layer architectures.

Despite the simplicity of these models, the visualizations offered valuable insights into how convolution and attention mechanisms influenced feature extraction. This evaluation provided a foundation for understanding zero-shot classification, domain generalization, and inductive biases, informing future exploration of more complex architectures that can better leverage both local and global information to enhance performance.

4. Results

Dataset	ViT	VGG16	CLIP
CIFAR 10	95.38%	83.88%	85.0%
SVHN	57.85%	80.21%	16.98%
PACS - Photo trained	47.62%	41.48%	69.43%
PACS - Art trained	36.14%	15.31%	89.65%
PACS - Cartoon trained	31.19%	44.3%	67.50%
PACS - Sketch trained	56.66%	55.16%	70.09%

Table 1. Comparison of domain generalization and response to covariate shifts in discriminative (ViT, VGG16) and contrastive (CLIP) models, evaluated on IID datasets (CIFAR-10) and OOD datasets (PACS and SVHN).

The results in Table 1 highlight the models' ability to generalize under both covariate and semantic shifts. ViT performs best in IID settings (CIFAR-10 with 95.38%) due to its global context-capturing ability. However, it struggles with the PACS dataset, which exhibits covariate shifts, and with SVHN, which represents a semantic shift. For example, ViT achieves only 36.14% on PACS Art, indicating a struggle with adapting to distinct styles that require more focus on local details. VGG16, with its strong locality bias, performs well on datasets where local patterns dominate, such as CIFAR-10 (83.88%) and SVHN (80.21%). However, this reliance limits its performance in generalized domains with diverse styles, as seen in PACS Art (15.31%). CLIP, due to its contrastive training, balances global and local context, allowing it to adapt effectively to both covariate and semantic shifts. Its higher performance across PACS domains (e.g., 69.65% on PACS Art) and reasonable performance on SVHN (16.98%) make it the most robust for handling OOD conditions.

The results in Tables 2 and 3 provide insights into the inductive biases of ViT, VGG16, and CLIP, particularly when dealing with visual transformations such as shape emphasis, texture dominance, color shifts, noise, and scrambling.

In Table 2, we see the models' responses to different visual biases—shape, texture, and color. ViT shows a strong shape bias due to its ability to capture global context effectively, achieving 9.38% on the COCO Silhouettes dataset. However, it struggles with datasets emphasizing texture, such as CIFAR-10 Styled (23.40%), due to its lack of locality bias. VGG16, with its locality bias, performs well on texture-biased datasets like CIFAR-10 Styled (38.70%), as it can leverage local details effectively. However, VGG16's performance drops significantly when global structure becomes crucial, as in COCO Silhouettes (4.46%), indicating difficulty in capturing overall object shapes. CLIP, with its balanced reliance on both local and global cues due to contrastive training, performs well across both shape and texture-biased datasets, as seen in its results on COCO (49.56%) and CIFAR-10 Styled (60.0%), highlighting its versatility and adaptability to various OOD conditions.

Dataset	ViT	VGG16	CLIP
COCO	60.27%	41.29%	49.56%
COCO silhouettes	9.38%	4.46%	20%
CIFAR 10	95.38%	83.88%	85.0%
CIFAR 10 Colored	57.70%	41.34%	43.98%
CIFAR 10 Styled	23.40%	38.70%	60.0%

Table 2. This table exposes the effects of inherent inductive biases of the discriminative and contrastive models by comparing its accuracies on datasets (COCO and CIFAR 10) before and after applying a transformation (color, shape, texture) on them

Dataset	ViT	VGG16	CLIP
CIFAR 10	95.38%	83.88%	85.0%
CIFAR 10 Noisy	57.18%	38.42%	39.46%
CIFAR 10 Scrambled	14.77%	13.65%	18.01%

Table 3. This table evaluates the effects of locality biases on the discriminative and contrastive models by comparing its accuracies on dataset (CIFAR 10) before and after applying some transformations on the image, by trying to capture how well, which model captures local and global dependencies

Table 3 extends this analysis by evaluating the robustness of the models under noisy and scrambled transformations, which further stress-test their inductive biases. On CIFAR-10 Noisy, ViT achieves 57.18%, indicating a moderate drop from the clean CIFAR-10 accuracy (95.38%). This drop suggests that while ViT is resilient to some level of added noise, its reliance on global context is partially affected by noise interfering with the overall structure. VGG16, on the other hand, shows a significant decrease in performance (38.42%) when noise is introduced, indicating that its reliance on precise local features makes it more susceptible to disturbances. CLIP achieves 39.46%, which, while lower than ViT, shows that it retains a reasonable level of recognition, benefiting from its balanced approach to feature extraction.

On CIFAR-10 Scrambled, all models show significant drops, but CLIP still performs slightly better (18.01%) compared to ViT (14.77%) and VGG16 (13.65%). Scrambling disrupts the spatial relationships between features, which particularly affects VGG16's locality-based recognition and ViT's

ability to derive meaningful global context. CLIP's ability to balance between recognizing global and local features provides it with a slight advantage in handling scrambled inputs, although the performance remains low for all models.

Combining the observations from Tables 2 and 3, we can see how each model's inductive bias influences its robustness to different types of image transformations. ViT excels in recognizing shape-related features but struggles with local details, making it vulnerable to transformations like texture shifts and scrambling. VGG16 is effective for texture recognition but is highly sensitive to noise and struggles when global structure is needed, such as in shape emphasis or scrambled conditions. CLIP emerges as the most adaptable model, demonstrating a balance that allows it to handle diverse transformations, whether related to shape, texture, color, noise, or spatial structure, resulting in better performance in OOD conditions compared to the other models. This adaptability stems from CLIP's contrastive training, which allows it to understand multiple visual cues effectively, minimizing performance drops across different challenging scenarios.

Bias	ViT	VGG16	CLIP
Texture	0.245	0.4613	0.706
Shape	0.152	0.090	0.404
Color	0.573	0.493	0.518

Table 4. This table shows the inductive biases of the discriminative and contrastive models by comparing its accuracies on dataset (CIFAR 10 and COCO) before and after applying some color shifts, style transfers, segmentations on the images. It is discussed in more detail, in a visualization under the Discussions section.

The results in Table 4 show the inductive biases of ViT, VGG16, and CLIP, reflecting how each model relies on texture, shape, and color features during classification tasks. CLIP has the highest texture bias (0.706), indicating a strong reliance on textural information, likely due to its contrastive learning approach, which helps capture detailed local features. VGG16 also exhibits considerable texture bias (0.4613), consistent with its convolutional architecture that emphasizes local feature extraction. ViT, with the lowest texture bias (0.245), is more focused on capturing global representations rather than local details.

For shape bias, CLIP shows a balanced reliance (0.404), effectively using both global structure and local details. ViT has a lower shape bias (0.152), indicating a tendency to capture more abstract representations rather than focusing on object shapes. VGG16, with the lowest shape bias (0.090), relies primarily on local features, limiting its effectiveness for shape-based distinctions.

Regarding color bias, ViT exhibits the highest value (0.573), suggesting a stronger dependence on color, which could make it effective when color is key but vulnerable when color information is inconsistent. CLIP (0.518) and VGG16 (0.493) have moderate color biases, providing more robustness in scenarios where color is not a reliable feature.

Overall, CLIP exhibits a strong texture bias and a balanced reliance on shape and color, making it versatile and effective for handling both local patterns and global structures. VGG16 has a dominant texture bias, making it effective at recognizing local details but less capable of focusing on global shapes. ViT shows a strong color bias and lower reliance on texture and shape, which may limit its generalization in scenarios where color is not informative. These biases provide insights into how each model's architecture and training approach affect their feature extraction strategies and performance in different visual conditions.

5. Discussion

Figure 4. This is an example of a diffusion model trying to conditionally denoise noised images, given a prompt ('airplane' in this case) at different timesteps.

In Figure 4, we can observe how the diffusion process generates an airplane image at different stages of denoising, revealing the underlying mechanics of the model's ability to create and modify visual content. At earlier timesteps, there is a higher level of noise applied to the latent representation, which gives the model more flexibility to explore different possible structures. As the diffusion process progresses, the model gradually transforms the noise into a recognizable shape, successfully forming an airplane, which indicates its capability to capture and enhance specific visual features as noise is progressively removed.

As the timesteps progress, the amount of noise decreases significantly, which reduces the model's ability to diverge from the original representation. During these later timesteps, the model has less opportunity to introduce significant changes, resulting in outputs that remain closely tied to the initial latent structure in terms of shape, texture, and color. In these stages, the denoising process focuses more on refining details rather than introducing novel variations, making it challenging for the model to create something entirely different from the original latent features. This behavior emphasizes the model's tendency to retain the essential characteristics of the image when noise levels are minimal.

However, if we increase the number of inference steps while keeping the same timesteps, the model gains additional iterations to modify the latent representation, providing it with more opportunities to diverge from the original image's visual structure. With more steps, the model can generate an image that aligns more closely with the guided prompt, even if it requires deviating significantly from the original shape, texture, or color. Essentially, increasing the number of inference steps allows the model to further explore the latent space, enabling greater creative freedom and the ability to produce novel visual content that departs from the original image frame.

This dynamic nature of the diffusion process demonstrates the trade-off between maintaining fidelity to the original content and introducing new variations. At earlier timesteps, the high noise levels give the model greater flexibility to explore different possibilities, while at later timesteps, the reduced noise leads to a more faithful reconstruction of the initial image. By adjusting the number of inference steps, we can effectively control the level of transformation applied to the image—either staying close to the original or deviating to create something entirely new as per the guided prompt. This illustrates the powerful capability of diffusion models to balance between preserving an image's original features and generating creative outputs based on specific guidance.

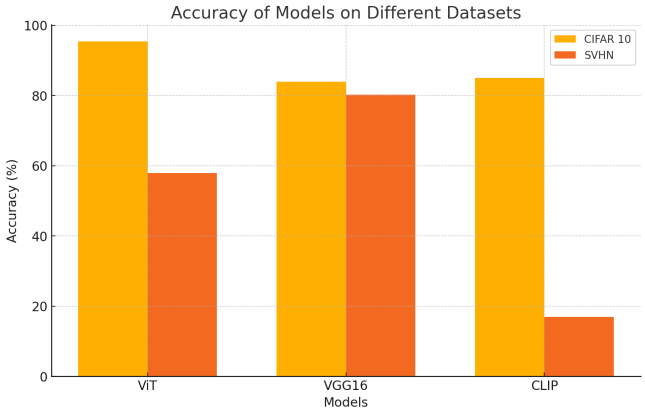

Figure 5. Comparison of model accuracies (ViT, VGG16, CLIP) on CIFAR-10 and SVHN datasets, highlighting significant differences in performance across datasets inside and outside domain.

Figure 5 presents a graphical comparison of model accuracies (ViT, VGG16, and CLIP) on CIFAR-10 and SVHN, highlighting the challenges in generalizing between in-domain (CIFAR-10) and out-of-domain (SVHN) data. This comparison emphasizes the difficulties faced by ViT and CLIP when moving from CIFAR-10 to the different semantic context of SVHN.

ViT, which performs well on CIFAR-10, shows a substantial drop in accuracy for SVHN, highlighting its limitation in extracting the localized features crucial for digit recognition. CLIP also struggles with SVHN, indicating that its balanced reliance on both global and local features is not sufficiently specialized for the specific requirements of this dataset. The significant performance drops of ViT and CLIP illustrate their difficulty in adapting from CIFAR-10 to a new domain where fine-grained local details are essential.

In contrast, VGG16 shows relatively consistent performance across both datasets. The convolutional layers of VGG16 capture spatial hierarchies and local patterns effectively, making it robust to shifts in visual domains. Its architecture's focus on localized feature extraction proves advantageous for SVHN, demonstrating the strength of convolutional models in scenarios where fine details are critical for accurate recognition.

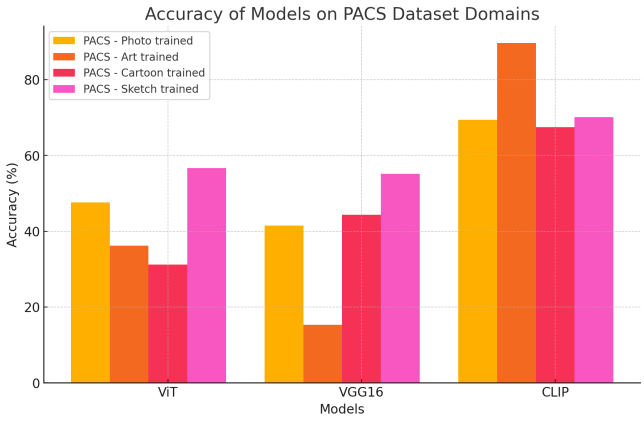

Figure 6. Accuracy of ViT, VGG16, and CLIP models finetuned on specific PACS dataset domains (Photo, Art, Cartoon, Sketch) and evaluated on others, illustrating model performance under domain shift conditions.

Figure 6 presents the accuracy of ViT, VGG16, and CLIP models when fine-tuned on specific PACS dataset domains (Photo, Art, Cartoon, Sketch) and evaluated on others, illustrating model performance under domain shift conditions. This comparison highlights the ability of each model to generalize across distinct visual styles, reflecting how their inherent inductive biases affect domain generalization.

CLIP consistently performs better across different domains, demonstrating its adaptability to a variety of visual styles. The model's balanced reliance on both global and local features allows it to effectively handle domain shifts between styles like Art and Cartoon. This flexibility is a direct outcome of CLIP's contrastive training, which provides robustness to diverse domains.

VGG16 shows moderate performance, benefiting from its convolutional layers that effectively capture localized fea-

tures. However, it struggles more than CLIP when faced with the more abstract styles like Art and Sketch, highlighting its limitations in generalizing beyond texture-heavy or locally distinct features. Despite this, VGG16 performs relatively well in domains that retain structural consistency with the training data.

ViT shows the lowest performance across the PACS domains, indicating challenges in adapting to domain shifts involving distinct styles. Its reliance on capturing global representations makes it less effective for abstract or stylized datasets where local details are important for recognition. The lower accuracy in these domains suggests that ViT's architectural approach is less suited for scenarios with significant style variation, where local features play a critical role.

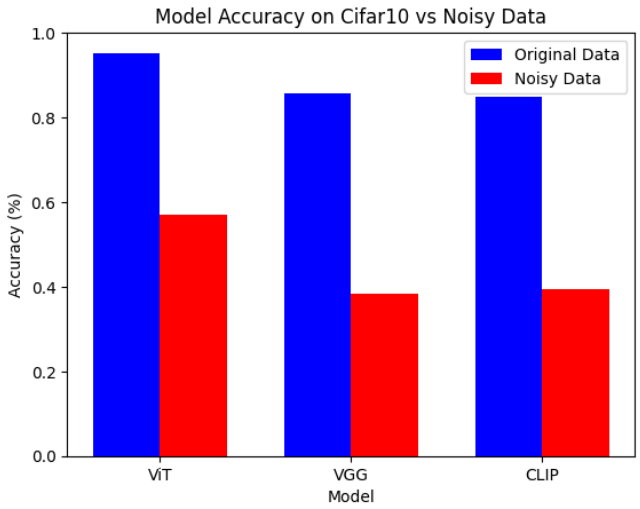

Figure 7. A comparison of the model accuracies on Cifar 10 vs noisy images of Cifar 10

Figures 7 and 8 compare the accuracy of ViT, VGG16, and CLIP models on CIFAR-10 and its modified versions, specifically noisy and scrambled images, to assess their robustness to data perturbations.

In Figure 7, both VGG16 and CLIP show reduced accuracy on noisy images but still retain some robustness due to their focus on local features. ViT experiences a larger drop, indicating its struggle with noise due to its reliance on global context rather than localized details.

Figure 8 shows that all models face significant drops in accuracy on scrambled images, with CLIP performing slightly better. VGG16 and ViT suffer considerably as scrambling disrupts spatial relationships, affecting both local and global feature extraction. CLIP's balanced feature reliance helps it retain some performance.

Overall, VGG16 and CLIP are more robust to noise, while CLIP slightly outperforms others on scrambled images. ViT is most affected in both scenarios due to its dependence on capturing intact global context. These results underscore each model's limitations and strengths under data perturbations.

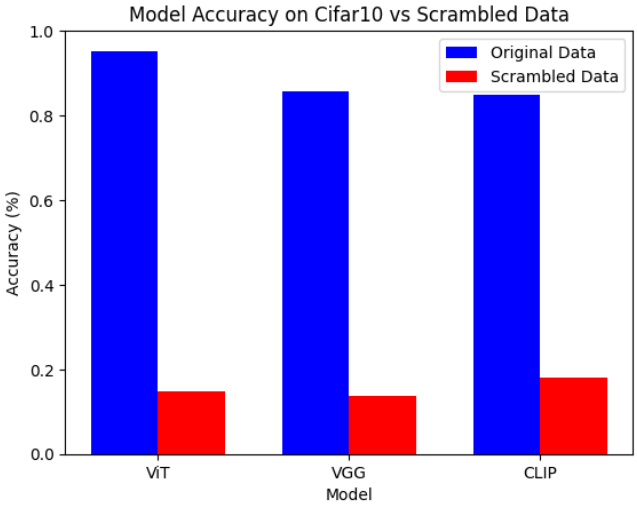

Figure 8. A comparison of the model accuracies on Cifar 10 vs scrambled images of Cifar 10

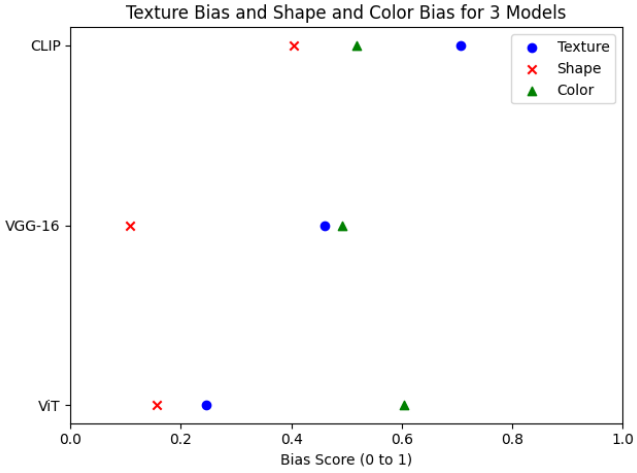

Figure 9. A comparison of the 3 models and their inductive biases after evaluation on rich datasets like CIFAR 10 and COCO.

The evaluation in Figure 9 compares the inductive biases of ViT, VGG16, and CLIP based on their reliance on texture, shape, and color after being tested on datasets like CIFAR-10 and COCO. The results illustrate their distinct preferences: VGG16, as a CNN, has a strong bias toward local dependencies, while ViTs prefer capturing global structures due to their reliance on self-attention. CLIP shows

a balanced approach, incorporating both local and global information, which enhances its versatility across different scenarios.

Interestingly, CLIP is more texture-biased than even VGG16 and has a higher shape bias than ViT. This blend of biases is likely due to CLIP's contrastive training, enabling it to capture both detailed textures and holistic shapes effectively. Despite having a transformer-based backbone like ViT, CLIP retains a significant texture bias, thanks to its training on diverse, real-world data where both textures and shapes are critical.

VGG16's strong texture bias aligns with its convolutional nature, excelling at extracting localized features but struggling with global structural understanding. ViT, with the lowest texture and shape biases, relies more on abstract, high-level features and may be less effective where detailed patterns or clear shapes are crucial.

In terms of color bias, all models exhibit similar scores, indicating that color is used but not relied upon heavily for classification. This shows that texture and shape play a more significant role for these models in challenging datasets like CIFAR-10 and COCO.

Overall, VGG16 relies heavily on local textures, ViT emphasizes global features, and CLIP balances between the two, contributing to its versatility and stronger performance in various tasks. This balanced bias helps CLIP effectively generalize across different datasets, highlighting each model's unique strengths and limitations in visual attribute understanding.

6. Conclusion

In this work, we evaluated the domain generalization capabilities and inductive biases of ViT, VGG16, and CLIP models. Systematic experiments on datasets with covariate and semantic shifts revealed that CLIP, with its balanced reliance on global and local features, performs best in out-of-distribution scenarios. ViTs excel in shape-biased tasks due to their focus on global context, while VGG16 performs well on texture-heavy datasets due to its locality bias. Our exploration of zero-shot classification with diffusion models showed limitations compared to CLIP in handling unseen data. Additionally, our study of single-layer models combining self-attention and depthwise convolution highlighted the potential benefits of hybrid models that effectively capture both local and global dependencies. These insights help pave the way for developing more robust and adaptive models for real-world applications.

Task	Done by
task 1	Adeen
task 2	Huraira
task 3	Adeen
task 4	Adeen except creation of shape bias dataset
task 5	Huraira
task 6	Huraira

7. Contributions

References

Clark, K., Jaini, P. (2024). Text-to-Image Diffusion Models are Zero-Shot Classifiers.

P. Gavrikov, J. Lukasik, S. Jung, R. Geirhos, B. Lamm, M. J. Mirza, M. Keuper, and J. Keuper. Are vision language models texture or shape biased and can we steer them?, 2024.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022.

J. Tian, Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira. Exploring covariate and concept shift for detection and calibration of out-of-distribution data, 2021.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey, 2024.

Hugging Face Diffusers Documentation: Hugging Face. (n.d.). Diffusers documentation. Retrieved from https://huggingface.co/docs/diffusers/index

Stable Diffusion Deep Dive (Notebook): fastai. (n.d.). Stable Diffusion Deep Dive. Retrieved from https://github.com/fastai/diffusion-nbs/blob/master/Stable%20Diffusion%20Deep%20Dive.ipynb