FREQUENCY DOMAIN IMAGE AUGMENTATION FOR DOMAIN GENERALIZED IMAGE CLASSIFICATION

by
SINA SALEH

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabanci University
July 2025

FREQUENCY DOMAIN IMAGE AUGMENTATION FOR DOMAIN GENERALIZED IMAGE CLASSIFICATION

Approved by:

Assoc. Prof. ÖZNUR TAŞTAN .....
(Thesis Supervisor)

Assoc. Prof. HÜSEYİN ÖZKAN .....

Prof. PINAR DUYGULU ŞAHİN .....

Date of Approval: JULY 17, 2025

ABSTRACT

FREQUENCY DOMAIN IMAGE AUGMENTATION FOR DOMAIN GENERALIZED IMAGE CLASSIFICATION

SINA SALEH

COMPUTER SCIENCE AND ENGINEERING M.Sc. THESIS, JULY 2025

Thesis Supervisor: Assoc. Prof. ÖZNUR TAŞTAN
Thesis Co-Supervisor: Prof. ERCHAN APTOULA

Keywords: Domain Generalization, Frequency Domain Augmentation, Domain Shift, Fast Fourier Transform

Domain Shift remains a major challenge in Domain Generalization (DG), where models trained on source domain(s) tend to perform poorly on unseen target domains. One effective approach to address this problem is the use of data augmentation techniques that synthetically enhance domain diversity. In this thesis, I introduce a frequency-domain augmentation method called Amplitude-Phase Augmentation (APA). APA operates by multiplying the amplitude components of source images with those from other domains in the frequency domain, while preserving the original phase information. This controlled mixing leads to the creation of cross-domain images that retain semantic structure but carry varied textural cues, increasing the robustness of models to distributional changes. I evaluate APA on two standard DG benchmarks: PACS and VLCS, using three diverse backbone architectures—ResNet-50, T2T-ViT-14, and DeiT-Small. APA is implemented on top of a standard Empirical Risk Minimization (ERM) framework and is also tested in conjunction with existing DG strategies. Extensive experiments show that APA improves generalization performance across both datasets and three backbones. Notably, APA achieves competitive results compared to strong baselines and recent augmentation-based methods on PACS dataset and superior results on VLCS across all three backbones. In addition to performance evaluations, I conduct detailed ablation studies on the amplitude mixing strategy and its effect on model robustness. These results demonstrate the practical effectiveness and adaptability of APA as a lightweight and domain-agnostic augmentation method for DG tasks. Code available at https://github.com/sina-nuel/APA

ÖZET

ALAN GENELLEŞTIRILMIŞ GÖRÜNTÜ SINIFLANDIRMASI IÇIN FREKANS ALANI GÖRÜNTÜ ARTTIRMA

SINA SALEH

BİLGİSAYAR BİLİMİ VE MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, TEMMUZ 2025

Tez Danışmanı: Doç. Dr. ÖZNUR TAŞTAN
Tez Eş Danışmanı: Prof. Dr. ERCHAN APTOULA

Anahtar Kelimeler: Alan genellemesi, Frekans alanı artırma, alan kaydırma, hızlı Fourier dönüşümü

Alan genellemesi (DG), modellerin kaynak alan(lar) üzerinde eğitildikten sonra hiç görmedikleri hedef alanlarda düşük performans sergilemesi nedeniyle bilgisayarla görmede hâlâ önemli bir sorundur. Bu sorunu aşmanın yollarından biri, kaynak alanlardaki veri çeşitliliğini sentetik olarak artıran görüntü artırma yöntemlerini kullanmaktır. Bu tezde, frekans alanı temelli Genlik–Faz Artırımı (Amplitude–Phase Augmentation, APA) adlı yeni bir yöntem önerilmektedir. APA, orijinal faz bilgisini korurken, kaynak görüntülerin genlik bileşenlerini diğer alanlardan elde edilen genliklerle karıştırarak yeni örnekler üretir. Bu sayede semantik içerik bozulmadan çeşitli dokusal ve frekans özellikleri taşıyan, alanlar arası zenginleştirilmiş görüntüler elde edilir ve modeller dağılım değişimlerine karşı daha dayanıklı hâle gelir. APA'yı değerlendirmek için iki yaygın DG benchmark'ı olan PACS ve VLCS üzerinde; ResNet-50, T2T-ViT-14 ve DeiT-Small olmak üzere üç farklı mimari kullanılarak deneyler gerçekleştirildi. Kapsamlı sonuçlar, APA'nın hem veri setlerinde hem de mimariler genelinde genelleme başarısını önemli ölçüde artırdığını gösteriyor. Özellikle, PACS'te güçlü temel yöntemlerle rekabetçi performans elde edilirken, VLCS'de üç mimaride de belirgin bir üstünlük sağlanmıştır. Buna ek olarak, genlik karıştırma stratejisinin model sağlamlığına katkısını değerlendirmek amacıyla ayrıntılı ablasyon çalışmaları yapıldı. Elde edilen bulgular, APA'nın DG görevlerinde alandan bağımsız ve pratik açıdan uygulanabilir bir artırma yöntemi olduğunu ortaya koymaktadır.Kod şu adreste mevcuttur: https://github.com/sina-nuel/APA

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Assoc. Prof. Öznur Taştan, and my co-advisor, Prof. Erchan Aptoula, for their invaluable guidance, encouragement, and continuous support throughout the course of this research. Their insights and expertise have been instrumental in shaping both the direction and the quality of this work.

I thank The Scientific and Technological Research Council of Turkiye (TUBITAK) for the provided computational support under Contract 121E452.

My heartfelt thanks go to my mother, father, and brother for their endless love, patience, and moral support. Their belief in me has always been a source of strength.

Finally, I am grateful to my friends for their encouragement, understanding, and companionship throughout this journey.

dedicated to the field

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background & Motivation

In the rapidly evolving field of machine learning, the capacity of a model to generalize beyond the data it has seen during training is critical for real-world deployment. In traditional machine learning, we usually assume that the train and test data both come from a same distribution but this assumption is not accurate in many real practices, where there is a huge difference between the conditions under which train data is collected and test data during deployment. This discrepancy, commonly referred to as domain shift (Quionero-Candela et al., 2008), can severely hinder a model's performance when it encounters new, unseen data that differs in distribution from the training data. Addressing this challenge has led to the development of a research paradigm known as Domain Generalization (DG) (Muandet et al., 2013).

DG refers to the situation where a model is trained on one or more source domains and is expected to perform well on an entirely unseen target domain that is not accessible during training. In contrast to traditional training setups where training, validation, and test datasets are all drawn from a single domain, DG tasks demand that the model effectively transfer its learned knowledge to new domains with different statistical properties. During regular training, the goal is often to ensure consistent performance across subsets (e.g., training, validation, and testing) of the same domain. However, DG poses a more complex and realistic challenge: achieving high generalization performance across domains that may vary significantly due to factors such as imaging devices, acquisition conditions, sensor types, environmental conditions, or demographic differences.

To clarify this concept, consider the application of machine learning to image classification using the PACS dataset (Li et al., 2017), which contains images of houses

in different styles—such as a cartoon drawing and a photograph of a house. Both images are easily recognizable to the human eye because they share common, semantically meaningful features like doors, windows, and yards, which define the concept of a house regardless of style or texture.

However, for a machine learning model, the differences in texture, background, color schemes, and artistic style between the cartoon and photo images present a challenge. A model trained only on photographs might rely on domain-specific features such as realistic textures or lighting patterns and thus struggle to correctly identify cartoon houses. Therefore, it is necessary for the model to learn to focus on domain-invariant features—like the shapes and spatial arrangements of doors, windows, and roofs—rather than superficial cues such as texture or style.

This ability to learn robust, shape-based features enables the model to generalize well across different domains and image styles. Figure 1.1 illustrates an example of two house images from the PACS dataset—one cartoon and one photo—highlighting how their common semantic features allow humans to recognize them as the same class despite their visual differences.

Solving this problem is not just of academic interest but is essential for building robust and dependable machine learning systems that can operate in diverse, real-world environments. For instance, in medical diagnosis, autonomous driving, remote sensing, and surveillance systems, it is often impractical or even impossible to anticipate all variations in future deployment scenarios. Therefore, models must be designed to generalize to new domains without explicit retraining or access to target domain data. This sets DG apart from Domain Adaptation (DA) (Ben-David et al., 2007), which typically assumes that a small sample of target domain data is available during training to fine-tune the model.

Over the years, a variety of approaches have been proposed to tackle the domain shift (Quionero-Candela et al., 2008) challenge. Some of the most prominent strategies include adversarial learning (Goodfellow et al., 2015), which aims to minimize the discrepancy between domains by fooling a domain discriminator; meta-learning (Thrun, 1998), which simulates domain shifts during training to prepare the model for unseen domains; and contrastive learning (Ganin et al., 2016), which encourages the model to learn representations that are invariant under certain transformations. However, the majority of these methods focus on feature extraction in the spatial domain and often struggle to fully overcome the limitations imposed by domain-specific noise or biases present in the visual appearance of input data.

In response to these limitations, a growing body of research has turned attention to

(a) (b)

Figure 1.1 Comparison of two house images, (a) is an image from cartoon domain and (b) is an image from photo domain from the PACS dataset. Despite differences in style, texture, and background, both images share common semantic features such as doors, windows, and triangle-shaped patterns as roofs, which enable human recognition.

the frequency domain as a complementary or alternative approach. One promising technique involves using the Fast Fourier Transform (FFT) (Cooley & Tukey, 1965) to decompose input images into their frequency components. This transformation enables a more nuanced understanding of the underlying structures within data by introducing amplitude and phase. By filtering or manipulating these frequency components, it is possible to suppress domain-specific variations and emphasize features that are more likely to generalize across domains.

Recent studies suggest that incorporating frequency-domain representations can significantly enhance the robustness and transferability of learned features. For instance, removing or down-weighting high-frequency noise that varies across domains can help prevent overfitting to domain-specific textures or artifacts. Additionally, frequency-based augmentation techniques can introduce novel training variations that simulate domain shift (Quionero-Candela et al., 2008), further preparing the model for deployment in diverse environments. Thus, frequency-domain methods offer a powerful and underexplored avenue for improving DG (Muandet et al., 2013).

In this thesis, I explored the potential of frequency-domain representations, specifically leveraging the FFT (Cooley & Tukey, 1965), to improve DG performance. I proposed an FFT-based framework that integrates frequency information into the feature learning pipeline, enabling the model to focus on domain-invariant characteristics and suppress irrelevant domain-specific cues. My approach is designed to operate solely on source domain data, making it well-suited for scenarios where target domain data is unavailable during training. Through comprehensive experiments on benchmark DG datasets, I evaluate the effectiveness of my method and compare it against existing techniques. My findings demonstrate that frequency-domain enhancements can lead to more robust and generalizable models, offering new insights into the role of frequency decomposition in mitigating domain shift.

1.2 Contributions

This thesis presents APA, a novel image augmentation technique rooted in the frequency domain, designed specifically to address the challenges of the domain shift (Quionero-Candela et al., 2008). The core motivation behind this method is to exploit the distinct and complementary roles of amplitude and phase components in the frequency domain to generate diverse and semantically meaningful variations of training data. My approach leverages the amplitude spectrum as a controllable factor to synthesize new data representations that are both realistic and domain-diverse, thereby enhancing the model's ability to generalize to previously unseen domains.

The primary contributions of this work can be summarized as follows:

I proposed a data augmentation framework that operates in the frequency domain by modifying the amplitude spectra of images while preserving the critical semantic content encoded in the phase. Unlike conventional augmentation techniques (Krizhevsky et al., 2012) that work in the spatial domain, APA enables the creation of spectrally altered yet structurally consistent samples by blending the amplitude components across different source domains. This strategy emphasizes learning domain-invariant features and suppresses domain-specific noise or artifacts, ultimately leading to better generalization across domains.

To better control the influence of augmentation during training, I introduce two new hyper-parameters that regulate when and how frequently the APA technique

is applied. The Augmentation Step Interval (S) defines how often augmentation is injected into the training process and the Augmentation Application Ratio (R) specifies the proportion of augmentation steps within each interval. This flexible control mechanism allowed me to balance learning from original and augmented data, and I observed that applying augmentation in later parts of each interval yields improved performance by allowing the model to first focus on stable learning before encountering more diverse representations.

In an extended version of APA, I introduced a refined augmentation strategy that incorporates attention mechanisms from Transformer-based models (Vaswani et al., 2017), such as ViTs (Dosovitskiy et al., 2021), into the frequency domain. The key insight motivating this enhancement is that not all spatial regions contribute equally to semantic understanding. By leveraging learned attention maps, I selectively emphasize or preserve the frequency components corresponding to regions that are most relevant for recognition. This attention-guided frequency modulation amplifies class-discriminative features while minimizing interference from background or uninformative regions, leading to more targeted and effective augmentations. The integration of attention into APA not only enhances model robustness but also aligns the augmentation process with the model's internal representation of salient visual cues.

I validated the effectiveness of the proposed APA method through extensive experiments on two widely-used domain generalization benchmarks: VLCS (Zhou et al., 2017) and PACS (Li et al., 2017). These datasets encompass a variety of domains such as photographs, art paintings, cartoons, and sketches. My experimental results demonstrate that APA improves performance across multiple target domains and achieves competitive results when compared to state-of-the-art DG techniques. The analysis further highlights APA's capacity to generate meaningful augmentations that contribute to improved robustness and generalization.

In summary, this thesis contributes a perspective to DG by incorporating frequency-domain into the training pipeline. Through the introduction of APA, I demonstrated that leveraging the frequency characteristics of image data provides a powerful mechanism for enhancing domain robustness. This work lays the foundation for future research at the intersection of spectral analysis and robust machine learning, and underscores the untapped potential of frequency-aware augmentations in improving model generalization across diverse and unseen domains.

2. RELATED WORK

As I explained before in chapter 1, domain shift (Quionero-Candela et al., 2008) is a critical challenge in computer vision, in which models trained on source domain(s) struggle to generalize to target domains due to distribution discrepancies (Zhou et al., 2022). Existing approaches typically address this issue through DA (Ben-David et al., 2007), which requires access to target data during training, or through DG (Muandet et al., 2013), which aims to learn transferable features from the source domain(s) with no access to the target domain(s). DG is generally more challenging due to the absence of target domain data and is a fast-growing area of research with various approaches proposed in the state-of-the-art studied by (Demirel et al., 2023).

Adversarial learning (Goodfellow et al., 2015; Ganin & Lempitsky, 2015; Li et al., 2019) is a widely adopted technique in DG, typically involving two competing components: a domain discriminator and a feature extractor. The core idea is to learn domain-invariant representations by having the feature extractor generate features that the domain discriminator cannot reliably distinguish across different domains. The discriminator, on the other hand, is trained to correctly classify the domain of each input, thereby setting up an adversarial objective. Through this mini-max game, the feature extractor is encouraged to produce features that are indistinguishable between source and target domains, effectively removing domain-specific information and preserving task-relevant patterns (Goodfellow et al., 2015). This process results in a shared feature space where samples from different domains are closely aligned, thus improving the model's ability to generalize to unseen target domains (Ganin et al., 2016).

A seminal contribution to this line of work was made by Ganin & Lempitsky (2015), who proposed a framework that combines unsupervised representation learning with adversarial training. They demonstrated how adversarial objectives could be leveraged to reduce discrepancies between feature distributions across domains, thereby enhancing the discriminative power of learned features for downstream prediction tasks. One of the key innovations in their approach was the introduction of the

gradient reversal layer (GRL), a mechanism that allows gradients from the domain discriminator to be reversed during backpropagation. This facilitates simultaneous optimization of the feature extractor to minimize task-specific loss while maximizing domain confusion, thereby ensuring that the learned features are both discriminative and domain-invariant.

More recently, Li et al. (2019) proposed FCN, an adversarial learning-based approach tailored for heterogeneous DG. Their method introduces a meta-learning framework that jointly trains a feature extractor and a critic network, where the critic serves as an adversary aiming to discriminate among domains based on extracted features. By using a meta-optimization strategy, the feature extractor learns to produce representations that not only minimize task loss but also confuse the critic across multiple source domains, encouraging stronger domain invariance. This approach extends traditional adversarial learning by incorporating meta-learning to better handle diverse domain shifts and has demonstrated improved generalization performance on several challenging benchmarks. The incorporation of a critic network in a meta-learning loop enables the model to dynamically adapt its feature representation to unseen target domains, making it a significant advancement over prior adversarial DG methods.

Overall, adversarial learning has emerged as a cornerstone technique in modern DA and DG frameworks, offering a principled and effective way to decouple task-relevant knowledge from domain-specific noise.

Meta-learning (Thrun, 1998; Li et al., 2018; Wang et al., 2025), often referred to as "learning to learn," (Thrun, 1998) is a powerful paradigm in DG that focuses on training models capable of quickly adapting to new, unseen domains using only prior experience from source domains. Unlike traditional machine learning approaches that train a model on a fixed distribution of data, meta-learning aims to expose the model to a variety of learning episodes during training, each mimicking a different domain shift. This enables the model to develop transferable knowledge and robust representations that can generalize effectively to novel domain distributions encountered at test time.

In the context of DG, meta-learning approaches typically simulate domain shifts by partitioning the available source domains into artificial training and testing tasks. During each iteration, a subset of source domains is treated as meta-train domains, while the remaining ones are considered as meta-test domains. The model is trained on the meta-train set and then evaluated and updated based on its performance on the meta-test set. This process encourages the learning algorithm to minimize the generalization error across tasks, ultimately leading to a model that is less sensitive

to domain-specific artifacts and better prepared for unseen domain distributions. This training paradigm not only enforces diversity in learning but also provides a mechanism for explicitly evaluating generalization ability during training.

A notable implementation of this idea is the model-agnostic meta-learning (MAML) algorithm, introduced by Li et al. (2018) for the purpose of DG. MAML optimizes for a set of model parameters that can be quickly fine-tuned to new tasks (or domains) with minimal data and updates. In the DG setting, this method was adapted to simulate domain shifts by creating meta-learning episodes, each involving randomly sampled meta-train and meta-test domain splits. Through this meta-training procedure, the model learns domain-invariant patterns and representations that generalize well across a spectrum of distributions, without requiring access to the target domain during training.

Building on this foundational work, subsequent meta-learning approaches for DG have explored more sophisticated task construction strategies, regularization techniques, and hierarchical learning structures. Some methods integrate adversarial or contrastive losses into the meta-learning loop, while others focus on explicitly modeling the inter-domain relationships to improve generalization. Finally, meta-learning has proven to be a highly effective and flexible strategy for enabling models to internalize generalizable knowledge, making it a central component in the development of robust DG algorithms. Recently Wang et al. (2025) proposed an innovative arithmetic meta-learning approach specifically aimed at DG Termed B-DMC. This method refines first-order meta-learning by not only matching gradients across different source domains, but by also driving model updates toward the centroid of domain-specific optimal parameters. The key insight is that aligning only the gradients is insufficient, as there are multiple possible descent directions; instead, B-DMC formulates an arithmetic combination of gradients, approximating the average optimum across domains. Extensive experiments on standard DG benchmarks show that this arithmetic-weighted strategy yields more balanced generalization, outperforming first-order meta-learning baselines. This contribution highlights the importance of parameter-space centering in addition to gradient alignment for achieving robust domain-invariant learning.

Augmentation methods (Krizhevsky et al., 2012; Guo et al., 2023; Demirel et al., 2023; Yang & Soatto, 2020; Zhou et al., 2021) are a widely adopted strategy in DG that aim to address the domain shift problem by artificially increasing data diversity during the training process. These methods operate under the premise that training a model on a broader and more varied distribution of inputs can help it learn more robust and domain-invariant features, which in turn improves its ability to generalize

to novel target domains. Traditional augmentation techniques (Krizhevsky et al., 2012) typically involve applying a series of input-space transformations, such as geometric operations (e.g., rotations, scaling, flipping), photometric adjustments (e.g., brightness, contrast, hue shifts), and spatial perturbations. These alterations serve to introduce synthetic variability into the training data, thereby mitigating the risk of overfitting to the specific visual characteristics of the source domains.

In recent years, augmentation strategies (Guo et al., 2023; Demirel et al., 2023; Yang & Soatto, 2020; Zhang et al., 2018), have grown significantly more sophisticated, incorporating not only input-level changes but also transformations in feature space and semantic representations. Some approaches, employ token-level or patch-level augmentation strategies, where parts of inputs (e.g., tokens, patches, or regions) are altered, replaced, or recombined to simulate cross-domain variability (Yun et al., 2019), while feature-space mixing methods combine features from different samples or domains to generate new, hybrid representations in the feature space (Zhang et al., 2018). These techniques are particularly effective in promoting invariance to superficial domain-specific cues and fostering a more generalized representation learning process.

Another influential approach that addresses the challenge of DG through a different lens is MixStyle by Zhou et al. (2021). In this work Rather than explicitly constructing pairs for contrastive training, MixStyle focuses on augmenting style statistics at the feature level to simulate domain shifts during training. The authors observe that style captured by channel-wise feature statistics such as mean and variance varies significantly across domains and is often a key factor in domain shift. MixStyle operates by randomly mixing the style statistics of feature maps between different images, effectively generating new feature-level domain variations within the same mini-batch.

This simple yet effective augmentation is integrated into existing architectures by inserting MixStyle layers after selected convolutional blocks. During training, the model is thus exposed to a continuously evolving spectrum of domain styles, forcing it to learn features that are more robust to such variations. Importantly, MixStyle does not require domain labels or access to target domain data, aligning well with the assumptions of DG. By training the model to be invariant to artificial style perturbations, MixStyle encourages the learning of semantic representations that generalize better across domains.

A compelling line of research has recently focused on frequency-based augmentation, where transformations are applied in the frequency domain rather than in the pixel or feature domain. For instance, Guo et al. (2023) proposed ALOFT, a method

that explicitly models the distribution of low-frequency components which tend to encode shape and structure and then uses these statistics to synthesize novel training samples. By preserving structural integrity while introducing frequency-based variability, their approach generates realistic yet diverse data that enhances the model's robustness to domain shifts. Similarly, the work of Yang & Soatto (2020) introduced the Fourier Domain Adaptation (FDA) method, which leverages the FFT (Cooley & Tukey, 1965) to manipulate amplitude components of images. Specifically, they proposed replacing a portion of the amplitude spectrum of a source image with that of another image belonging to the same class but from a different domain. This controlled frequency-based manipulation allows the model to focus on class-relevant features while discarding irrelevant domain-specific signals.

Augmentation-based approaches offer a flexible and powerful mechanism for improving generalization in DG settings. By simulating diverse forms of domain variability during training, these techniques help models learn to extract invariant patterns that are less sensitive to superficial shifts in appearance, structure, or distribution. As augmentation strategies continue to evolve, particularly through advances in frequency analysis and generative modeling, they are likely to remain a critical component of robust domain generalization pipelines.

Contrastive learning (Hadsell et al., 2006; Motiian et al., 2017) is a self-supervised representation learning paradigm that aims to learn meaningful embeddings by distinguishing between similar and dissimilar pairs of data points. The core principle involves pulling together embeddings of semantically similar samples while pushing apart those of dissimilar ones within the feature space (Hadsell et al., 2006). In the context of DG, contrastive learning is particularly effective because it encourages the model to focus on intrinsic semantic content rather than domain-specific artifacts. By enforcing consistency across semantically similar samples irrespective of their domain of origin, contrastive learning facilitates the development of more robust and transferable representations that generalize well to unseen domains.

The learning process typically involves constructing positive and negative pairs or triplets of samples. Positive pairs are composed of instances that share the same class label or represent different views (e.g., augmentations) of the same instance, while negative pairs come from different classes. The model is trained using a contrastive loss such as InfoNCE (van den Oord et al., 2018) or triplet loss (Schroff et al., 2015) which guides the embedding space to maintain this semantic structure while disregarding superficial domain-level variations.

An approach that leverages this principle is proposed in Motiian et al. (2017), where the authors introduced a Siamese network architecture (Bromley et al., 1993) cou-

pled with a contrastive semantic alignment loss for DG. Their method explicitly constructs pairs of samples from different domains but with the same class label as positive pairs, and pairs from different domains and different classes as negative ones. The contrastive loss then regulates the feature space such that intra-class samples, regardless of their domain origin, are pulled closer together, while inter-class samples are pushed further apart. This architecture ensures that class identity, rather than domain identity, drives the geometry of the learned representation space. As a result, the model becomes less sensitive to domain-specific cues and more attuned to task-relevant semantics.

The key advantage of contrastive learning in DG lies in its ability to impose a form of self-supervision that aligns the model's feature space according to semantic similarity while being agnostic to domain membership. This alignment plays a critical role in enabling generalization to novel, unseen domains, particularly in settings where domain-induced variations can otherwise overwhelm the learning signal.

Feature alignment (Muandet et al., 2013; Sun & Saenko, 2016; Gulrajani & Lopez-Paz, 2020; Wei et al., 2021) is another fundamental technique in DG that aims to explicitly reduce domain discrepancies by aligning the feature distributions of different domains, either through statistical measures or geometric transformations (Muandet et al., 2013). The underlying motivation is to encourage models to learn domain-invariant representations, such that features extracted from inputs across varying domains share similar distributions in the latent space. This approach helps mitigate the domain shift problem by ensuring that the model's decision boundaries remain stable when presented with data from unseen target domains.

Feature alignment can be carried out at different levels of statistical abstraction. Early methods primarily focused on aligning first-order statistics (e.g., means), but more recent and effective strategies go further by aligning second-order statistics such as covariances. A notable contribution in this space is the work by Sun & Saenko (2016), who proposed Deep CORAL (Correlation Alignment)—a method that aligns the second-order statistics of features extracted from multiple source domains. Their approach uses nonlinear transformations within deep neural networks to minimize the discrepancy in covariance matrices across domains. This encourages the model to learn a shared feature space that is agnostic to domain-specific structural differences.

Building on these ideas, Gulrajani & Lopez-Paz (2020) introduced a more generalized framework for feature alignment in DG. Their method penalizes both the mean and covariance differences across domain distributions using a well-defined loss term, thereby promoting tighter alignment across the entire statistical profile of domain-specific features. Importantly, their work also underscored the importance

of architectural and algorithmic choices in DG, demonstrating through an extensive empirical study that model selection plays a critical role in the effectiveness of feature alignment methods.

While many feature alignment techniques assume that all extracted features are equally relevant to the downstream task, more recent research has challenged this assumption. For instance, Wei et al. (2021) proposed a novel approach that separates features into two categories: task-relevant and task-irrelevant. Their insight was that aligning all features indiscriminately might dilute the effectiveness of the alignment process by giving equal importance to noisy or domain-specific features that do not contribute meaningfully to the classification task. To remedy this, they introduced a dual-objective framework consisting of both alignment and feature weighting. The alignment process was applied only to task-relevant features, which were identified using a Gradient-weighted Class Activation Mapping (Grad-CAM) technique introduced by (Selvaraju et al., 2017). Grad-CAM allows the model to generate spatial attention maps that highlight the most discriminative regions within Convolutional Neural Network (CNN) (LeCun et al., 1998) feature maps, effectively guiding the model to focus on the most informative aspects of the input data.

This enhanced alignment strategy ensures that the model prioritizes features that are not only invariant across domains but also critical for classification. Such selective alignment has proven to be more robust than uniform alignment methods, especially in scenarios with complex or high-dimensional domain variations. More generally, recent advancements in this area have explored class-conditional alignment, adversarial alignment, and alignment in latent or semantic spaces to further improve the precision and effectiveness of feature-based generalization strategies.

Finally, feature alignment remains an important part of DG research, offering a principled means of addressing distribution shifts. As techniques become more refined and incorporate attention, weighting, or domain decomposition mechanisms, they continue to improve the reliability of models operating in open and dynamic environments.

Transfer learning (Pan & Yang, 2010; Hu et al., 2022; Lee et al., 2024; Li et al., 2023) represents a crucial strategy within the DG landscape, focusing on the reapplication of knowledge gained from a source task or domain to improve performance on a related but distinct target task or domain. Unlike traditional learning paradigms that assume identical distributions between training and test data, transfer learning acknowledges the existence of domain shifts and seeks to bridge this gap by transferring useful representations, parameters, or inductive biases from previously learned models (Pan & Yang, 2010). This approach is particularly valuable in DG,

where labeled data from the target domain are unavailable during training. By leveraging pretrained models often trained on large, diverse datasets, transfer learning provides a computationally efficient and empirically effective alternative to training from scratch, significantly reducing the need for extensive domain-specific labeled data.

At the core of transfer learning is the idea that many tasks share underlying structures and features, especially in high-dimensional domains such as computer vision or natural language processing. As a result, a model trained to extract informative features in one setting can offer strong priors for downstream tasks (e.g., object detection or scene understanding) even in domains with different visual characteristics. In the context of DG, this reuse of pretrained knowledge can help mitigate overfitting to source domains and improve robustness to unseen distributions by anchoring learning in broadly applicable feature representations.

Recent advancements in this area have explored parameter-efficient fine-tuning (PEFT) methods like LoRA (Hu et al., 2022) to adapt large pretrained models without the need for full retraining. A compelling example of this is found in the work of Lee et al. Lee et al. (2024), who proposed a PEFT framework that enables adaptation of large-scale foundation models to new tasks within domain-generalized settings. Specifically, they introduced a novel architecture called Mixture of Adapters (MoA), where instead of fine-tuning the entire model, a series of lightweight adapters are trained and inserted into the network. These adapters vary in capacity quantified by their internal rank and are governed by a learnable routing mechanism that dynamically selects the most appropriate adapter based on the characteristics of the input and the task. This modular design allows the model to generalize across diverse domains while maintaining high parameter efficiency, making it both scalable and adaptive to new challenges without significant computational cost.

A parallel line of work by Li et al. (2023) also embraced the idea of specialization through modularity, introducing a sparse expert routing approach for DG. Their method can be viewed as a rule-based transfer learning framework, where each expert module functions like a logical reasoning unit that fires based on specific feature conditions. For instance, in visual recognition tasks, one expert might activate upon detecting large ears and curved tusks interpreted as rules for recognizing elephants. This interpretable, symbolic-inspired logic mimics the "if/then" reasoning process and provides a different flavor of transfer learning, where task-specific knowledge is encoded into discrete expert functions that the model selectively activates based on the input. Such expert-based models allow for flexible task transfer and can be fine-tuned or expanded with minimal interference between experts, thus helping to avoid

negative transfer and preserve generalization performance across diverse domains.

Together, these methods highlight the versatility of transfer learning within DG, where the challenge is not just transferring information, but doing so selectively and efficiently across varying tasks and domains. As models continue to grow in scale and complexity, approaches like adapter-based fine-tuning (Houlsby et al., 2019), expert routing (Shazeer et al., 2017), and sparsely activated sub-networks (Lepikhin et al., 2021) are likely to play a central role in scalable and interpretable DG systems.

3. PROPOSED METHOD

This chapter outlines the methodology developed and employed throughout the course of this thesis. It presents a chronological overview of the conceptual and technical evolution of the method, beginning with the initial inspiration and culminating in the final proposed approach.

3.1 Frequency-Based Domain-Specific Features Separation

3.1.1 Motivation

Deep learning models, particularly CNNs (LeCun et al., 1998), have achieved remarkable success across a variety of computer vision tasks, including classification, segmentation, and detection. However, when it comes to DG (Muandet et al., 2013) these models often struggle due to their inherent biases and limitations in architectural design.

One key architectural limitation arises from CNNs' strong texture bias. Prior studies have demonstrated that CNNs tend to rely excessively on local texture cues rather than capturing holistic object representations, which can result in fragile generalization when textures change across domains (Choi et al., 2023). This is largely because CNNs process images through localized convolutional filters, which inherently focus on small, spatially constrained regions of the input. While hierarchical layers in CNNs aim to progressively aggregate local information into more global representations, the inductive bias of locality limits their capacity to effectively model long-range dependencies and global structures in images (Baker et al., 2018). As a

result, CNNs are often sensitive to superficial appearance changes and can fail when faced with novel domain-specific distortions or shifts.

Meanwhile, the emergence of Vision Transformers (ViTs) (Vaswani et al., 2017) has introduced a new paradigm for visual representation learning. Unlike CNNs, ViTs employ self-attention mechanisms to model global dependencies across an image, treating it as a sequence of patches. This architecture grants ViTs a natural advantage in capturing long-range contextual relationships, which are often crucial for recognizing global object configurations. However, despite their strength in modeling global patterns, ViTs may still struggle to preserve or exploit fine-grained local details when global attention is over-prioritized. Hence, even ViTs can benefit from integrating complementary modalities such as frequency domain representations that explicitly encode local and global information in a structured, disentangled manner.

To overcome these limitations, researchers have explored alternative representations of image data most notably, the frequency domain. The frequency domain offers a powerful perspective by transforming spatial pixel information into frequency components via the FFT (Cooley & Tukey, 1965). This transformation decomposes an image into sinusoidal basis functions, allowing the representation of both low-frequency components, which capture coarse, global image structure (e.g., shape, layout), and high-frequency components, which encode fine-grained, local details (e.g., edges, textures) (Cai et al., 2021). This dual representation enables models to capture a more comprehensive understanding of image content, which is crucial for DG, where robustness to both global and local domain shifts is needed.

Frequency-based approaches have shown promise in recent DG and DA studies. For instance, researchers have proposed modifying or replacing only certain parts of the amplitude spectrum (typically the low-frequency components) in the FFT representation to synthesize new training data that mimic domain shifts (Zheng et al., 2022). While this selective amplitude manipulation can introduce valuable cross-domain variability, it also poses risks: focusing exclusively on low-frequency or high-frequency cues may result in the loss of important complementary information and lead to suboptimal learning of domain-invariant features.

In this work, I proposed an image augmentation technique called APA for DG, which operates in the frequency domain. My approach is motivated by the insight that different domains often share similar global and local patterns, even if they vary in visual appearance. By manipulating images in the frequency domain, APA allows us to construct synthetic images that blend spectral characteristics from multiple source domains, thus simulating domain shifts in a controlled and meaningful way.

The APA method is designed as a DG technique that leverages the frequency domain properties of images to produce more robust and diverse training samples. Specifically, APA works by taking two images that belong to the same semantic class but originate from different source domains. These domains might vary in lighting, texture, resolution, or imaging conditions, which are common sources of distribution shift in real-world datasets. By transforming both images into the frequency domain using the FFT, APA extracts their respective amplitude and phase spectra. It then performs a controlled element-wise multiplication of their amplitude components, while preserving the phase spectrum of one of the original images.

The motivation for this specific choice of manipulation stems from the distinct roles that amplitude and phase play in the visual composition of an image. The amplitude spectrum encodes the strength and intensity of different spatial frequency components essentially capturing texture, fine details, and structural patterns such as edges or repetitive textures. In contrast, the phase spectrum retains information about the spatial arrangement and semantic content of the image such as the shape and positioning of objects within a scene (Yang & Soatto, 2020). These two components together determine how an image is perceived. However, phase has been shown to be significantly more important than amplitude in preserving the recognizable structure of objects. For instance, if you reconstruct an image using only the phase spectrum and a uniform or neutral amplitude spectrum, the resulting image still retains the general layout and identifiable features, although it may appear washed out or less sharp. On the other hand, reconstructing an image using only the amplitude and zero phase yields a ghost-like or abstract pattern, lacking semantic meaning but still containing some texture clues from the original image.

To illustrate this point, Figure 3.1 demonstrates the visual effects of isolating and recombining phase and amplitude spectra. The figure includes the original image alongside its amplitude and phase representations. One reconstructed version shows the effect of combining the original phase with a uniform amplitude emphasizing that even without the original strength of frequency components, the semantic and structural integrity of the image remains largely intact. Another reconstruction shows the image built from amplitude only, with the phase set to zero. This version appears abstract or ambiguous, with localized textures and edges but no coherent structure. This experiment highlights the critical importance of phase in defining image semantics, while amplitude plays a complementary role in encoding style and appearance.

APA leverages this understanding by mixing amplitude components across domains while retaining the original phase. By doing so, it introduces controlled stylistic vari-

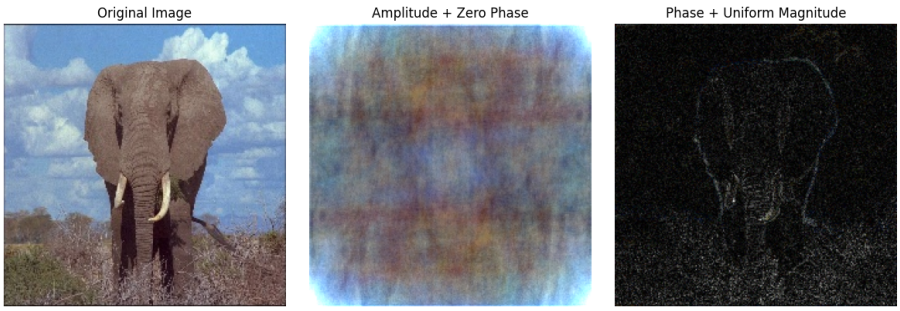

Figure 3.1 Showing each frequency component effect solely to better understand them

ation such as textures, lighting variations, and structural cues from other domains without altering the fundamental semantic content of the image. This controlled augmentation produces synthetic images that lie between domains and exhibit a blend of styles from multiple sources. These intermediate samples effectively act as bridges across domains, helping the model generalize better by learning features that are not tightly coupled to domain-specific textures or lighting conditions. In practice, such augmentation contributes to the learning of domain-invariant representations, which are crucial for models that must perform well across unseen target domains where distributional characteristics differ from training data and introduces a frequency-domain perspective to data augmentation that is grounded in the fundamental properties of image composition. By separating and recombining amplitude and phase and understanding their respective roles, APA offers a principled approach to creating semantically consistent but stylistically diverse training data.

APA does not restrict the augmentation to a particular frequency range. Instead, it applies a comprehensive transformation that better maintains the multi-scale structure of the image. Furthermore, this augmentation can be seamlessly integrated into the training pipeline of both CNNs and ViTs, enabling the models to learn from a richer spectrum of variations without any architectural changes.

I hypothesized that training on images augmented via APA exposes the model to a broader and more diverse distribution of domain-induced variations, thereby improving its generalization capability. Specifically, APA simulates domain shifts within the training phase, helping the model learn to ignore domain-specific noise and focus instead on semantically meaningful, task-relevant features. The augmented data serve as a proxy for unseen domains, improving the model's ability to operate in open-world settings.

To validate my approach, I evaluated APA on two widely used DG benchmarks: VLCS (Zhou et al., 2017) and PACS (Li et al., 2017). These datasets encompass

diverse domains with varying styles and content, providing a strong testbed for DG methods. Empirical results demonstrate that APA leads to competitive performance compared to baseline and state-of-the-art methods. My findings confirm that frequency-domain augmentation especially through principled combinations of amplitude and phase components can enhanced a model's robustness to domain variability.

Initially, the idea was to combine images in the frequency domain by multiplying both their phase and amplitude components. According to the convolution theorem (Bracewell, 1999), multiplication in the frequency domain corresponds to convolution in the spatial domain. In other words, if two images are multiplied in the frequency domain, the resulting image in the spatial domain is equivalent to the convolution of the original images by using the second image as the kernel.

Convolution in the spatial domain is a fundamental operation in image processing, used to apply filters such as edge detectors, blurring, or sharpening. It includes moving a matrix as a filter over another image or matrix and calculating the weighted sum of the overlapping values at each cell. Mathematically, for an image $I(x, y)$ and a kernel $K(u, v)$, the convolution is defined as:

$$(I * K)(x, y) = \sum_u \sum_v I(x - u, y - v) \cdot K(u, v). \tag{3.1}$$

where:

- $I(x, y)$: The input image intensity at spatial coordinates (x, y).
- $K(u, v)$: The convolution kernel (or filter) value at position (u, v).
- $(I * K)(x, y)$: The result of convolving the image I with the kernel K at position (x, y).
- $\sum_u \sum_v$: Double summation over the kernel's spatial dimensions (i.e., iterating over all values of u and v that define the size of the kernel).
- $I(x-u, y-v)$: The image value at position $(x-u, y-v)$, corresponding to the shifted position under the kernel.
- $\cdot$: Multiplication of the image value with the corresponding kernel value.

This operation effectively combines local neighborhood information, and its behavior depends on the choice of the kernel. So I was curious what happens if I use an image with same content but different style as the kernel and what happens to the resulted

augmented image.

However, as discussed earlier, modifying the phase component was found to be unnecessary. Moreover, directly multiplying the amplitude spectra does not result in visually meaningful augmented images and therefore requires additional modifications, which will be detailed in the following sections.

3.1.2 Technical Approach

The first step of the method involves applying the FFT on the input images I and I', converting them to the frequency domain components. FFT allows for transforming spatial domain information (pixel-based representations) into that of the frequency domain, which breaks down the image into its frequency components,phase and amplitude.

For each channel c (Red, Green, Blue) of a color image, I compute its frequency domain representation using the discrete 2D Fourier transform:

$$F_c(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_c(x,y)\, e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.2}$$

where:

- $I_c(x,y)$ is the intensity value of channel c at spatial location (x,y),
- $F_c(u,v)$ is the complex-valued frequency domain representation at frequency coordinates (u,v),
- $M \times N$ is the spatial resolution of the image,
- $u \in \{0,1,\ldots,M-1\}$ and $v \in \{0,1,\ldots,N-1\}$ are discrete frequency indices along the horizontal and vertical directions, respectively,
- i is the imaginary unit ($i^2 = -1$).

The frequency representation $F_c(u,v)$ can be expressed in polar form as:

$$F_c(u,v) = A_c(u,v)\, e^{i\Phi_c(u,v)}, \quad \text{where } A_c(u,v) = |F_c(u,v)|, \quad \Phi_c(u,v) = \arg(F_c(u,v)). \tag{3.3}$$

where:

- $|F_c|$ denotes the modulus (magnitude) of the complex number,
- $\arg(F_c)$ denotes the phase angle (argument) of the complex number.

Similarly, for the second image I', I obtain its amplitude A'_c and phase Φ'_c. Once the images are transformed into the frequency domain, their amplitudes are multiplied. Convolution in the spatial domain is equivalent to multiplying both phase and amplitude in the frequency space. However, in the APA method, only amplitudes are used. This operation allows me to alter specific frequency components like amplitude here, thereby changing the image's characteristics. By multiplying the amplitudes and taking the square root, new augmented images are generated that share key semantic properties with the original images and introduce variations that help the model generalize better to unseen domains.

To construct a new image, a transformed amplitude $\tilde{A}_c(u,v)$ is defined by the element-wise geometric mean of the amplitudes of two images:

$$\tilde{A}_c(u,v) = \sqrt{A_c(u,v) \cdot A'_c(u,v)} + \varepsilon, \tag{3.4}$$

where:

- $A_c(u,v)$ and $A'_c(u,v)$ are the amplitude spectra of the two images for channel c,
- ε is a small positive constant added to prevent division by zero or instability,
- $\tilde{A}_c(u,v)$ is the mixed amplitude to be used in the synthesis process.

Since the amplitude represents the texture or style of an image, and the phase carries the structural content, we retain the original phase components during recombination. The synthetic frequency components are then computed as:

$$\tilde{F}_c(u,v) = \tilde{A}_c(u,v) \cdot e^{i\Phi_c(u,v)}, \quad \tilde{F}'_c(u,v) = \tilde{A}_c(u,v) \cdot e^{i\Phi'_c(u,v)}, \tag{3.5}$$

where:

- $\Phi_c(u,v)$ and $\Phi'_c(u,v)$ are the phase components of the two images,

- $\tilde{F}_c(u,v)$ and $\tilde{F}'_c(u,v)$ are the resulting frequency-domain representations combining shared amplitude with distinct phases.

To obtain the corresponding spatial-domain synthetic images, the Inverse Discrete Fourier Transform (IDFT) is applied:

$$\tilde{I}_c(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \tilde{F}_c(u,v) \cdot e^{2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.6}$$

$$\tilde{I}'_c(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \tilde{F}'_c(u,v) \cdot e^{2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.7}$$

where:

- $\tilde{I}_c(x,y)$ and $\tilde{I}'_c(x,y)$ are the resulting synthetic images in the spatial domain for channel c,
- M and N are the height and width of the image,
- i is the imaginary unit.

The aim of this process is to create two synthetic images that retain the structural properties of their respective input classes while combining their domains' texture and style. After the generation of the augmented images, the model is trained using both the original and the augmented images.

By exposing the model to a broader range of image variations, the proposed approach aims to learn transferable features across domains. The augmented images not only increase the diversity of the training set but also encourage the model to focus on the underlying structure of the data, rather than relying on domain-specific features that could hinder generalization. This improves performance when the model is applied to unseen target domains. The architecture of my model is depicted in Fig.3.2.

In addition to the standard hyper-parameters for model training, there are two additional hyper-parameters in the augmentation phase:

- Augmentation Step Interval (S): The number of training steps after which augmentation is periodically applied within. For example, if S is set to 200, augmentation is applied every 200 steps of the training.

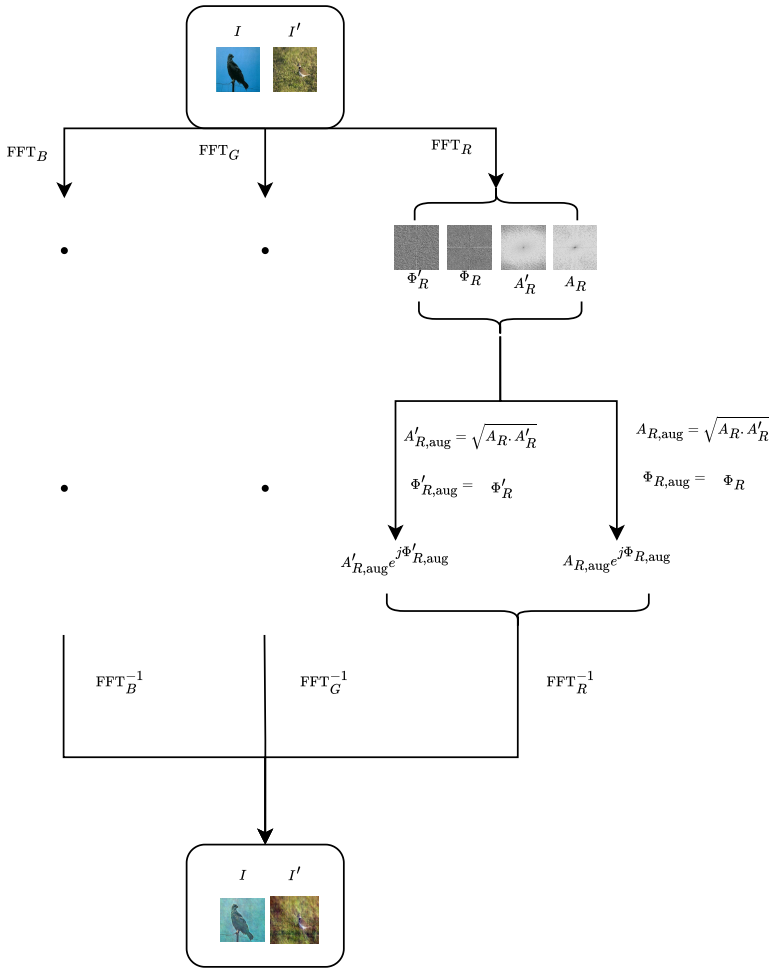

Figure 3.2 Detailed augmentation procedure in the APA module. For simplicity, detailed operations are shown only on channel R, the two big dots showing two levels of the operations as same as shown for channel R on other two channels. Both I and I' images were used to create two augmented images. A_R, A'_R ,ϕ_R and ϕ'_R are the amplitudes and phases of the I and I' images in R channel respectively. $A_{R,\mathrm{aug}}$, $A'_{R,\mathrm{aug}}$,$\phi_{R,\mathrm{aug}}$ and $\phi'_{R,\mathrm{aug}}$ are the amplitudes and phases of the augmented images in R channel respectively. For each augmented image, the phase of one image was used to preserve its main structure, while the style of the domains was mixed using the amplitudes. The dot product of the amplitudes is element-wise. FFT_R, FFT_G and FFT_B refer to FFT operations on the three channels, and FFT_R^{-1}, FFT_G^{-1} and FFT_B^{-1} are their respective inverse FFT operations.

- Augmentation Application Ratio (R): The proportion of steps within each interval to which augmentation is applied. For example, if R is 0.1 and S is 200, in 20 steps out of every 200 steps, I apply the augmentation and in the remaining 180 steps I train the model with input images without augmentation. I found that training first with the original input images and then synthesized images led to better performance, as discussed in the results section.

My APA method offers several advantages. First, it allows me to generate diverse samples from existing data without requiring additional labeled examples from the target domain. Second, modifying the frequency components ensures that the introduced variability remains general, reducing the likelihood of capturing domain-specific artifacts such as color and texture. Additionally, working in the frequency domain provides greater flexibility for mathematical operations. For example, instead of multiplying amplitudes directly, we take the square root of their product to smooth the combination process. Fig. 3.3 exemplifies the difference between augmented images using APA with and without taking the square root and convolution itself. In natural images, low-frequency components which usually correspond to smooth intensity variations, typically have large amplitudes. In contrast, high-frequency components, which capture edges, textures, and fine details, generally have much smaller amplitudes. Experimental observations on all images from the PACS dataset (Li et al., 2017) show that, on average, 46% of the high-frequency components have amplitudes less than one, compared to only around 1% for low-frequency components. When the square root operation is applied across the frequency spectrum, it amplifies values less than one. This selective amplification causes a noticeable portion of the high-frequency components to become more pronounced, while leaving the low-frequency components largely unaffected. As a result, the blurring caused by high-frequency attenuation which acts as a low-pass filter is counteracted (Kostkova et al., 2020). Since blurring suppresses high-frequency details that are crucial for sharpness and clarity (Gonzalez & Woods, 2007), enhancing these components can effectively restore image sharpness and reduce blurriness. Thus, applying the square root to frequency amplitudes serves as a controlled method for enhancing image details and mitigating high-frequency loss due to blurring.

3.2 Boosted Frequency-Based Domain-Specific Features Separation

(a)

(b)

(c)

(d)

(e)

Figure 3.3 Comparing APA and other convolution-based augmentations: (a) and (b) are input images; (c) augmented image using APA and taking (a) as the base image by applying square root on amplitude multiplication; (d) augmented image using APA and taking (a) as the base image without applying square root on amplitude multiplication; and (e) augmented image by multiplying both phases and amplitudes.

In this enhanced approach, I proposed a refined augmentation technique that builds upon the previously introduced frequency-based method by integrating the representational advantages of attention mechanisms from transformer models. By selectively emphasizing regions that are deemed important by a ViT (Vaswani et al., 2017) model we can potentially amplify class-discriminative features in the frequency domain. This section details my attempt to leverage attention maps (Vaswani et al., 2017) as spatial importance indicators and use them as a dynamic weighting factor during frequency-based image augmentation.

3.2.1 Motivation

The principal motivation for this hybrid method is that attention maps inherently contain spatial information about where the model focuses its semantic understanding. These maps are not random, they result from contextual interactions captured by the transformer layers and can serve as a soft prior for enhancing the frequency characteristics of informative regions. By interpreting the attention map as a grayscale intensity function, I treat high-attention regions as spatial zones of interest that often exhibit sharper spatial variations or localized features.

Computing the frequency-domain representation of these attention maps enabled me to capture underlying structural patterns associated with important semantic regions. High-attention areas, which tend to vary more sharply, contribute significant energy to the high-frequency components of the Fourier spectrum, while low-attention regions, often smoother and less informative, primarily influence the low-frequency components.

For instance, consider an image depicting a dog, where the model strongly focuses on the ear due to its distinctive geometry and textural cues. The resulting attention map will exhibit a bright region corresponding to the ear. When transformed into the frequency domain, this attention map emphasizes frequency components associated with that region's spatial structure. By fusing this with the image's original amplitude spectrum, I create an augmented image that retains the dominant textures and contours of semantically relevant regions, effectively boosting their representation in the frequency domain.

This method acts as a frequency-sensitive filtering process, where regions with higher semantic relevance identified by the transformer receive greater weight during spectral fusion. Consequently, the resulting images are not only more representative of

inter-domain variability but also maintain structural fidelity to the original features deemed most useful for classification. Moreover, this approach reduces the salience of less informative or distracting background elements, effectively guiding the model's attention during training toward features with higher discriminative value.

This boosted frequency-based feature separation technique introduces a new mechanism to integrate the strengths of attention-based interpretability with the efficiency and robustness of frequency-domain augmentation. By modulating the amplitude spectrum using transformer derived attention maps whose spatial variation encodes semantic importance the model is better equipped to learn spatial-frequency representations aligned with its internal attention mechanisms, fostering improved generalization and domain adaptability.

3.2.2 Technical Approach

The attention mechanism alone, while semantically meaningful, lacked the structural fidelity required to maintain low-level visual details in augmented images. Therefore, I investigated a more nuanced integration one that combines the frequency-domain representations of both the original images and their corresponding attention maps. Specifically, my goal was to incorporate attention as a weighting component during the amplitude modulation phase of the FFT-based image fusion process.

As introduced earlier, I decompose images I and I' into their frequency components by using the FFT. This decomposition separates the images into amplitude and phase components denoted as $A_c(u,v)$, $A'_c(u,v)$, $\phi_{\text{att}(u,v)}$ and $\phi'_{\text{att}(u,v)}$ respectively as defined by Equations 3.2 and 3.3. In the original method, I combined the amplitudes of two input images to generate a synthesized image in the frequency domain. In this improved method, I introduce a third component: the attention-weighted amplitude derived from the image's Transformer-based attention map.

To implement this, I begin by inputting two images denoted as I and I' into a pretrained Transformer model to extract their respective attention maps $att(I)$ and $att(I')$. I used the last attention map layer in both DeiT-Small (Touvron et al., 2021) and T2T-ViT-14 (Yuan et al., 2021) backbones. These attention maps, often represented as multi-head outputs or averaged attention heatmaps, are then converted into normalized grayscale images, referred to as I_{att} and I'_{att}. These grayscale attention maps serve as proxies for region-level importance, where brighter pixels indicate stronger attention. I subsequently apply the FFT to these maps to extract

their frequency components, obtaining both amplitudes and phases:

The discrete Fourier transforms of the attention maps are denoted as:

$$F_{\text{att}}(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_{\text{att}}(x,y)\, e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.8}$$

$$F'_{\text{att}}(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I'_{\text{att}}(x,y)\, e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.9}$$

where:

- $I_{\text{att}}(x,y)$ and $I'_{\text{att}}(x,y)$ are the intensity value the attention maps of the I and I' images at spatial location (x,y),
- $F_{\text{att}}(u,v)$ and $F'_{\text{att}}(u,v)$ are the complex-valued frequency domain representations of the attention maps of the I and I' images at frequency coordinates (u,v),
- $M \times N$ is the spatial resolution of the attention maps,
- $u \in \{0,1,\ldots,M-1\}$ and $v \in \{0,1,\ldots,N-1\}$ are discrete frequency indices along the horizontal and vertical directions, respectively,
- i is the imaginary unit ($i^2 = -1$).

Each complex Fourier coefficient can be decomposed into amplitude and phase components as:

$$F_{\text{att}(u,v)} = A_{\text{att}(u,v)}\, e^{i\phi_{\text{att}(u,v)}}, \quad \text{where } A_{\text{att}(u,v)} = |F_{\text{att}(u,v)}|, \quad \phi_{\text{att}(u,v)} = \arg(F_{\text{att}(u,v)}) \tag{3.10}$$

$$F'_{\text{att}(u,v)} = A'_{\text{att}(u,v)}\, e^{i\phi'_{\text{att}(u,v)}}, \quad \text{where } A'_{\text{att}(u,v)} = |F'_{\text{att}(u,v)}|, \quad \phi'_{\text{att}(u,v)} = \arg(F'_{\text{att}(u,v)}) \tag{3.11}$$

where:

- $|F_{\text{att}(u,v)}|$ and $|F'_{\text{att}(u,v)}|$ denote the modulus (magnitude) of the $F_{\text{att}(u,v)}$ and $F'_{\text{att}(u,v)}$,
- $\arg(F_{\text{att}(u,v)})$ and $\arg(F'_{\text{att}(u,v)})$ denotes the phase angle (argument) of the $F_{\text{att}(u,v)}$ and $F'_{\text{att}(u,v)}$.

I compute augmented amplitudes by combining the original and attention-derived amplitudes. This reinforces frequency components associated with regions of high attention:

(3.12)
$$\tilde{A}_c(u,v) = \sqrt[3]{A_c(u,v) \cdot A'_c(u,v) \cdot A_{\text{att}(u,v)}} + \varepsilon, \quad \tilde{A}'_c(u,v) = \sqrt[3]{A_c(u,v) \cdot A'_c(u,v) \cdot A'_{\text{att}(u,v)}} + \varepsilon$$

where:

- A_c, A'_c are the amplitude spectra of I_c and I'_c respectively,
- A_{att}, A'_{att} are the amplitudes from the attention maps,
- ε is a small constant added to prevent division by zero or numerical instability.

Using the original phase components $\phi_c(u,v)$ and $\phi'_c(u,v)$, the augmented frequency representations are reconstructed as:

$$\tilde{F}_c(u,v) = \tilde{A}_c(u,v)\, e^{i\phi_c(u,v)}, \quad \tilde{F}'_c(u,v) = \tilde{A}'_c(u,v)\, e^{i\phi'_c(u,v)} \tag{3.13}$$

The corresponding spatial-domain images are then obtained using the inverse Fourier transform:

$$\tilde{I}_c(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \tilde{F}_c(u,v) \cdot e^{2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.14}$$

$$\tilde{I}'_c(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \tilde{F}'_c(u,v) \cdot e^{2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)}, \tag{3.15}$$

where:

- $\tilde{I}_c(x, y)$ and $\tilde{I}'_c(x, y)$ are the resulting synthetic images in the spatial domain for channel c,
- M and N are the height and width of the image,
- i is the imaginary unit.

4. EXPERIMENTS

4.1 Experiments

In this section, I describe the experimental setup used to evaluate the effectiveness of the proposed APA method for DG (Muandet et al., 2013). My evaluation is conducted on two widely adopted benchmark datasets: VLCS (Zhou et al., 2017) and PACS (Li et al., 2017). These datasets have been extensively used in the literature for assessing DG techniques due to their inherent domain diversity and challenging nature. My goal is to rigorously assess how well APA generalizes to unseen domains when trained on a subset of available source domains.

4.1.1 VLCS Dataset

The VLCS dataset (Zhou et al., 2017) is a standard benchmark for domain generalization and is composed of images drawn from four heterogeneous domains: PASCAL VOC 2007 (V) (Everingham & et al., 2010), LabelMe (L) (Russell & et al., 2008), Caltech 101 (C) (Fei-Fei & et al., 2003), and SUN09 (S) (Xiao & et al., 2010). Each domain has distinct characteristics in terms of background clutter, lighting conditions, scene complexity, and image resolution. For example, the VOC images are mostly object-centric with cluttered backgrounds, while the Caltech images are relatively clean and captured in controlled environments. The SUN09 images, in contrast, contain natural scenes with diverse environmental settings, and LabelMe includes user-labeled web images with varying quality and context.

The dataset consists of five object categories that are shared across all domains:

bird, *car*, *chair*, *dog*, and *person*. Despite having the same set of semantic labels, these classes are visually quite different across domains due to the domain-specific imaging conditions and annotation sources. Such variability makes VLCS a suitable testbed for assessing the generalization capabilities of models.

For my experiments, I followed the widely used leave-one-domain-out evaluation protocol. Under this setting, for each experiment, I selected one domain as the target domain and used the remaining three as source domains for training. This ensures that the model is evaluated on a domain that it has not seen during training, simulating a real-world generalization scenario. I repeated this process for each of the four domains in the dataset, resulting in four distinct train-test splits. The classification performance was measured using average accuracy on the held-out domains.

4.1.2 PACS Dataset

In addition to VLCS, I evaluated my approach on the PACS dataset (Li et al., 2017), another benchmark specifically designed to challenge DG methods. PACS comprises four visually distinct domains: *Photo (P)*, *Art Painting (A)*, *Cartoon (C)*, and *Sketch (S)*. These domains represent a wide spectrum of visual styles, ranging from realistic photographs to highly abstract and artistic depictions.

PACS contains images belonging to seven object categories: *dog*, *elephant*, *giraffe*, *guitar*, *horse*, *house*, and *person*. Unlike VLCS, which focuses more on natural scenes and varied backgrounds, PACS emphasizes domain shifts related to drawing styles, line thickness, texture absence or exaggeration, and semantic abstraction. For instance, while a horse in the Photo domain appears with realistic texture and shading, its representation in the Sketch domain might be reduced to a few lines capturing only the silhouette.

Similar to my experiments on VLCS, I applied the leave-one-domain-out evaluation protocol for PACS. In each experiment, one domain was designated as the target (test) domain, and the model was trained on the remaining three source domains. This protocol was repeated across all four domains, and the classification performance on the unseen domain was recorded. Average classification accuracy was used as the evaluation metric, enabling fair comparison with prior DG methods in the literature.

4.1.3 Implementation Details

In all experiments, images were resized to a fixed resolution of 224×224 to maintain consistency with standard deep learning backbones. The APA augmentation was applied during training by manipulating the amplitude spectra of image pairs sampled from the same class but different domains, while preserving the original phase spectrum to maintain semantic content. To ensure robustness and reproducibility, each experiment was repeated three times with different random seeds. For baselines, I compared APA against both standard training without augmentation baseline and state-of-the-art domain generalization methods.

APA was implemented in PyTorch utilizing CNN (LeCun et al., 1998) based architecture such as ResNet-50 (He et al., 2016) and ViT (Vaswani et al., 2017) backbones such as DeiT-Small (Touvron et al., 2021) and T2T-ViT-14 (Yuan et al., 2021) as the classifiers. The proposed augmentation was applied during training to generate additional domain-variant samples. The model was trained with a total of 6000 epochs fo each backbone. The experiments were conducted on a server equipped with two NVIDIA RTX 3090 GPUs, each with 24GB of VRAM. You can find the hyper-parameters for each backbone in the provided table 4.1.

Table 4.1 Hyper-parameters used for APA training across different backbones and datasets.

Hyperparameter	ResNet-50 (PACS/VLCS)	T2T-ViT (PACS/VLCS)	DeiT-Small (VLCS)
batch_size	8	8	16
class_balanced	True	True	True
lr	3e-5	1e-5	1e-5
weight_decay	0.0001	0.001	0.0
S	200	200	200
R	0.1	0.4	0.4

Figure 4.1 provides a visual overview of sample images from both VLCS and PACS datasets. These examples highlight the diversity and complexity of the domain shifts present in each dataset.

4.1.4 Results and Discussion

To validate the effectiveness of the proposed APA method, I conducted extensive experimental evaluations comparing APA against a diverse set of competitive DG methods. These baseline methods were carefully selected based on their strong

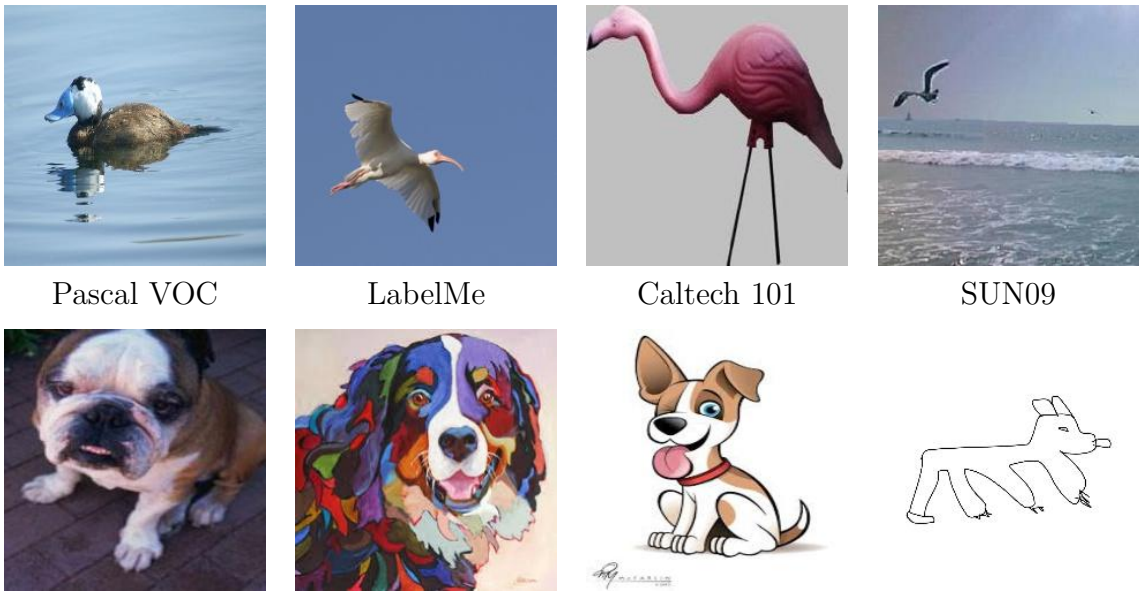

Figure 4.1 Examples from the VLCS and PACS datasets from the four available domains for the two classes, bird (top row-VLCS) and dog (bottom row-PACS).

Table 4.2 Classification accuracy (%) on PACS and VLCS using ResNet-50 backbone.

Method	PACS	VLCS
Baseline	84.5	77.5
Mixup (Yan et al., 2019)	84.6	77.4
MLDG (Li et al., 2018)	84.9	77.2
DANN (Ganin et al., 2016)	83.6	78.6
SagNet (Nam et al., 2020)	86.3	77.8
ARM (Zhang et al., 2020)	85.1	77.6
SWAD (Cha et al., 2021)	**88.1**	79.1
GMDG (Tan et al., 2024)	85.6	79.2
ADRMX (Demirel et al., 2023)	85.3	78.5
APA (mine)	85.7	**80.3**

reported performance in the literature, wide adoption, and compatibility with standard deep learning backbones. By choosing both classical and recent approaches, I ensured a comprehensive and fair evaluation. Furthermore, to eliminate confounding factors, all models including baselines and APA were implemented using the same training configurations and backbone architectures.

Table 4.3 Classification accuracy (%) on PACS and VLCS using T2T-ViT-14 and DeiT-Small backbones.

Backbone	Method	PACS	VLCS
T2T-ViT-14	Baseline	86.8	78.9
	TFS-ViT (Noori et al., 2024)	89.0	80.0
	SDViT (Sultana et al., 2022)	88.0	79.5
	APA (mine)	88.5	81.7
	Boosted APA (mine)	**89.8**	**81.8**
DeiT-Small	Baseline	84.9	78.8
	TFS-ViT	87.3	80.2
	SDViT	**88.3**	79.2
	APA (mine)	87.1	81.4
	Boosted APA (mine)	87.9	**81.7**

To assess the robustness of APA across architecture types, I experimented with both CNN and ViT based backbones. This dual setup allowed us to evaluate APA not only on traditional CNN structures but also on modern attention-based architectures, which are gaining traction in vision tasks.

The baseline methods include well-established DG algorithms such as Mixup (Yan et al., 2019), which performs linear interpolation in the input or feature space; MLDG (Li et al., 2018), a meta-learning-based approach designed to simulate domain shift during training; and DANN (Ganin et al., 2016), a domain-adversarial training method that learns domain-invariant features via gradient reversal. Additionally, I included more recent and competitive methods such as GMDG (Tan et al., 2024), which leverages group-wise feature alignment. SagNet (Nam et al., 2020) separates style and content representations to improve generalization, while ARM (Zhang et al., 2020) enhances robustness through adaptive risk minimization. SWAD (Cha et al., 2021) builds on stochastic weight averaging to stabilize training and improve out-of-domain performance. For ViT comparisons, TFS-ViT was studied (Noori et al., 2024), a transformer-specific strategy that employs task-specific feature sampling. I also included SDViT (Sultana et al., 2022), which integrates structural domain-specific priors into ViT training for better generalization. For a better understanding, a comparison of these methods is provided in Table 4.4.

Table 4.4 Summary of Compared Domain Generalization Methods

Method	**Description**
Mixup (Yan et al., 2019)	Performs linear interpolation of images and/or their features to generate synthetic training samples that help regularize the model.
MLDG (Li et al., 2018)	Meta-learning strategy that simulates domain shift during training to improve generalization to unseen domains.
DANN (Ganin et al., 2016)	Domain-Adversarial Neural Network that promotes domain-invariant feature learning via gradient reversal.
SagNet (Nam et al., 2020)	Separates style and content representations in the feature space to prevent overfitting to superficial statistics.
ARM (Zhang et al., 2020)	Adaptive Risk Minimization framework that learns robust predictors by minimizing risk under varying domain conditions.
SWAD (Cha et al., 2021)	Stochastic Weight Averaging in Domain Generalization that stabilizes training and improves out-of-domain performance.
GMDG (Tan et al., 2024)	Group-wise Meta Domain Generalization using feature alignment at the group level to improve robustness across domains.
TFS-ViT (Noori et al., 2024)	Task-specific Feature Sampling method tailored for ViT architectures to enhance discriminative learning across domains.
SDViT (Sultana et al., 2022)	Introduces structural domain-specific priors into ViT training to promote better generalization in structured data scenarios.

4.1.4.1 Results on CNN-based Backbones

Table 4.2 summarizes the average classification accuracy across all target domains for each method. Using ResNet-50 (He et al., 2016) as the CNN backbone, APA outperformed all of the classical and recent baselines on the VLCS dataset. Specifically, APA attained a significant improvement over MLDG and DANN and compared to Mixup, APA demonstrated a clear advantage, underscoring the benefits of selectively manipulating the amplitude spectrum while preserving semantic content through the phase.

On the PACS dataset, APA achieved competitive performance, closely matching or surpassing other methods in several domain splits. The results confirmed APA's ability to generalize across diverse styles without requiring style-specific feature tuning or domain labels during inference. The gains observed on PACS are particularly meaningful given the substantial variation in visual appearance between domains (e.g., the abstract nature of sketches compared to the realism in photos), which tends to degrade the effectiveness of shallow data augmentations or generic domain-invariant learning methods.

4.1.4.2 Results on Transformer-based Backbones

To further explore the adaptability of APA, I extended my experiments to transformer-based vision models, including T2T-ViT-14 and DeiT-Small, both of which represent compact yet powerful transformer architectures optimized for visual tasks. These experiments evaluated APA's capacity to function effectively in a non-convolutional feature extraction setting, which has a fundamentally different mechanism for spatial and contextual reasoning.

APA was benchmarked against DG methods designed specifically for transformer architectures. On the VLCS dataset, APA outperformed both TFS-ViT and SDViT on all target domains, establishing itself as a strong frequency-based augmentation strategy even in transformer settings. This is particularly noteworthy, as transformer models often require specialized training schemes or architectural modifications to achieve strong generalization. APA, in contrast, required no changes to the backbone and integrated seamlessly with the existing training pipeline, providing a lightweight and generalizable solution.

For the PACS dataset, although APA was not the best but it had competitive results

on this dataset too. One reason behind this is that the domain shift in this dataset is very bold and make it harder for the model to generalize well.

The empirical findings across both CNN and ViT backbones highlight several key insights. First, APA's success on VLCS suggests that frequency-domain manipulation is particularly effective in datasets where structural and background variations dominate the domain shift.

APA's effectiveness in transformer models underscores the generality of the method. Despite the differences in how CNNs and ViTs process images, both benefit from frequency-based augmentation, suggesting that APA operates on a level orthogonal to the architectural design. Unlike methods that rely on architectural bias, APA provides a plug-and-play solution that enhances model robustness through carefully crafted spectral transformations.

For the Boosted version you can see that I got a nice increase in accuracy for PACS dataset under both backbones, but for the VLCS it was not that much, it shows that this boosting method worked on a dataset with more difficult domain shift.

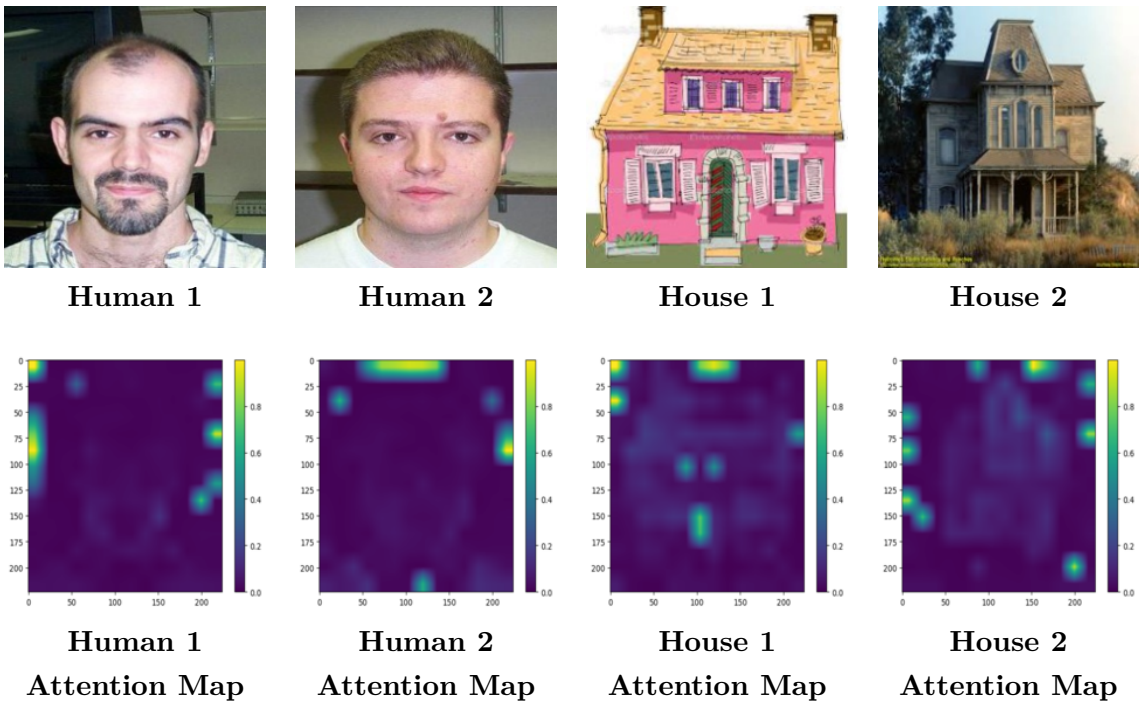

Figure 4.2 Examples from the PACS dataset showing two classes—human and houses (top row)—and their corresponding attention maps (bottom row). The attention maps highlight the generic and structurally important parts of each object (e.g., face contours, house outlines).

As shown in the attention maps in Figure 4.2, key object structures such as face contours or building shapes are consistently emphasized as highlighted areas in the middle of these images. These attention maps implicitly capture the semantic essence

Backbone	Dataset	Source Domains	Target Domain	Accuracy (%)
ResNet-50	PACS	P, C, S	A	87.6
ResNet-50	PACS	P, A, S	C	82.4
ResNet-50	PACS	A, C, S	P	97.1
ResNet-50	PACS	P, A, C	S	75.6
ResNet-50	VLCS	V, L, S	C	97.3
ResNet-50	VLCS	V, C, S	L	68.1
ResNet-50	VLCS	V, L, C	S	76.0
ResNet-50	VLCS	L, C, S	V	79.8

Table 4.5 Domain generalization results on PACS and VLCS datasets using ResNet-50 as the backbone.

Backbone	Dataset	Source Domains	Target Domain	Accuracy (%)
T2T-ViT-14	PACS	P, C, S	A	91.5
T2T-ViT-14	PACS	P, A, S	C	81.3
T2T-ViT-14	PACS	A, C, S	P	97.7
T2T-ViT-14	PACS	P, A, C	S	83.6
T2T-ViT-14	VLCS	V, L, S	C	98.1
T2T-ViT-14	VLCS	V, C, S	L	67.1
T2T-ViT-14	VLCS	V, L, C	S	79.6
T2T-ViT-14	VLCS	L, C, S	V	82.1

Table 4.6 APA Accuracy on T2T-ViT-14 as the backbone, PACS and VLCS as the datasets separately on each domain

of the object. When transferred to the frequency domain using FFT, both the original image and the attention map exhibit shared frequency components related to these dominant structures. By using the attention map as a weighting mechanism in the frequency domain, specifically by multiplying its amplitude with that of the original image, we can selectively amplify these semantically important and commonly shared frequency components. This enhances domain-relevant information while suppressing irrelevant variations, ultimately improving generalization across domains.

In tables 4.7, 4.6, 4.9, and 4.8 the results can be seen separately for each domain on both T2T-ViT-14 and DeiT-Small backbones and PACS and VLCS datasets.

The results in Table 4.10 present the class-wise accuracy of APA and Boosted APA on the PACS dataset using different backbones. Overall, Boosted APA consistently improves upon the baseline APA method across most object classes. For instance, with the T2T-ViT-14 backbone, Boosted APA increases the accuracy for the *Elephant* class from 85.8% to 90.0%, and for the *Person* class from 86.1% to 91.2%. Similarly, with DeiT-Small, improvements are observed in classes such as *Giraffe* (from 87.4% to 89.2%) and *Horse* (from 78.4% to 83.9%). In ResNet50 the performance

Table 4.7 APA Accuracy on DeiT-Small as the backbone, PACS and VLCS as the datasets separately on each domain

Backbone	Dataset	Source Domains	Target Domain	Accuracy (%)
DeiT-Small	PACS	P, C, S	A	89.6
DeiT-Small	PACS	P, A, S	C	78.1
DeiT-Small	PACS	A, C, S	P	96.9
DeiT-Small	PACS	P, A, C	S	77.1
DeiT-Small	VLCS	V, L, S	C	97.6
DeiT-Small	VLCS	V, C, S	L	67.9
DeiT-Small	VLCS	V, L, C	S	79.1
DeiT-Small	VLCS	L, C, S	V	81.0

Table 4.8 Boosted APA Accuracy on DeiT-Small as the backbone, PACS and VLCS as the datasets separately on each domain

Backbone	Dataset	Source Domains	Target Domain	Accuracy (%)
DeiT-Small	PACS	P, C, S	A	90.3
DeiT-Small	PACS	P, A, S	C	83.4
DeiT-Small	PACS	A, C, S	P	99.0
DeiT-Small	PACS	P, A, C	S	78.8
DeiT-Small	VLCS	V, L, S	C	97.7
DeiT-Small	VLCS	V, C, S	L	68.2
DeiT-Small	VLCS	V, L, C	S	79.3
DeiT-Small	VLCS	L, C, S	V	81.4

Table 4.9 Boosted APA Accuracy on T2T-ViT-14 as the backbone, PACS and VLCS as the datasets separately on each domain

Backbone	Dataset	Source Domains	Target Domain	Accuracy (%)
T2T-ViT-14	PACS	P, C, S	A	92.2
T2T-ViT-14	PACS	P, A, S	C	83.6
T2T-ViT-14	PACS	A, C, S	P	98.9
T2T-ViT-14	PACS	P, A, C	S	84.4
T2T-ViT-14	VLCS	V, L, S	C	98.2
T2T-ViT-14	VLCS	V, C, S	L	68.9
T2T-ViT-14	VLCS	V, L, C	S	78.4
T2T-ViT-14	VLCS	L, C, S	V	81.6

is generally lower, emphasizing the advantage of transformer-based architectures for this task.

Table 4.11 shows the class-wise accuracy on the VLCS dataset, where a similar pattern can be observed. Boosted APA consistently outperforms APA across several categories, although the performance gain is more moderate compared to PACS. For example, using DeiT-Small, the accuracy for the *Person* class increases from 74.9% to 76.2%, and with T2T-ViT-14, the *Dog* class improves from 62.8% to 66.2%. While

Backbone	Method	Dog	Elephant	Giraffe	Guitar	Horse	House	Person
T2t-ViT-14	APA	84.6	85.8	86.7	96.6	86.3	97.4	86.1
T2t-ViT-14	Boosted APA	78.5	90.0	86.4	97.1	84.4	97.0	91.2
DeiT-Small	APA	83.5	84.8	87.4	92.6	78.4	91.1	87.9
DeiT-Small	Boosted APA	84.2	84.4	89.2	92.9	83.9	92.9	86.9
ResNet50	APA	76.7	84.5	87.1	93.5	79.8	88.9	79.1

Table 4.10 Class-wise accuracy of the APA on PACS dataset with different backbones

Backbone	Method	Bird	Car	Chair	Dog	Person
T2t-ViT-14	APA	80.8	88.7	69.5	62.8	74.3
T2t-ViT-14	Boosted APA	79.4	87.6	71.5	66.2	75.3
DeiT-Small	APA	80.6	89.7	69.6	61.3	74.9
DeiT-Small	Boosted APA	79.9	87.4	69.7	62.3	76.2
ResNet50	APA	68.0	85.1	71.6	57.6	74.1

Table 4.11 Class-wise accuracy of the APA on VLCS dataset with different backbones

ResNet50 occasionally performs competitively on specific classes such as *Chair* and *Person*, its overall accuracy remains less consistent.

4.1.5 Visual Results

In this section I want to provide some visual results to give a better insight about this work. First I want to provide a visual comparison between the FDA (Yang & Soatto, 2020) and my approach.

In figure 4.3 you can see the comparison between two frequency-based augmentation methods, FDA(Yang & Soatto, 2020) and APA. As the FDA method was implemented on the GTA V (Richter et al., 2016) dataset, for the comparison I tried my method on this dataset too and you can see as I consider all range of the amplitudes including both high-frequency and low-frequency components it gives better augmented images with better details in the edges. In this comparison I used the second images from the VLCS (Zhou et al., 2017) dataset having two different conditions, in the first one the weather is cloudy and in the second one the image was captured at night, you can see the effect of the weather on the new augmented images. In the first row corresponding to the night one, you can see the augmented images has a darker effect and because the first image has sunny weather, the resulted image is like a sunset weather. In the second row again the base image has a sunny weather but the second image has a cloudy weather, you can see in the

Figure 4.3 Augmenting images from GTA V dataset with images from VLCS with two different conditions using FDA and APA method.

resulted image we have the effect of the cloudy weather comparing to the first sunny image.

The next visualization is about some samples of augmented images from the PACS and VLCS datasets in all four domains using a class, the dog class is used for the PACS and the chair class is used for the VLCS examples.

4.1.6 Ablation Study

To better understand the contribution and sensitivity of the proposed APA method, I conducted a detailed ablation study focusing on two critical hyper-parameters: the *augmentation step interval* and the *augmentation application ratio*. These hyper-parameters govern when and how frequently the APA augmentation is applied during training. While the augmentation step interval determines how often augmentation phases occur (i.e., the temporal frequency of augmentation), the augmentation application ratio controls how many of the steps within each interval are augmented. Together, these parameters form the basis of APA's temporal scheduling mechanism.

4.1.6.1 Motivation for Ablation

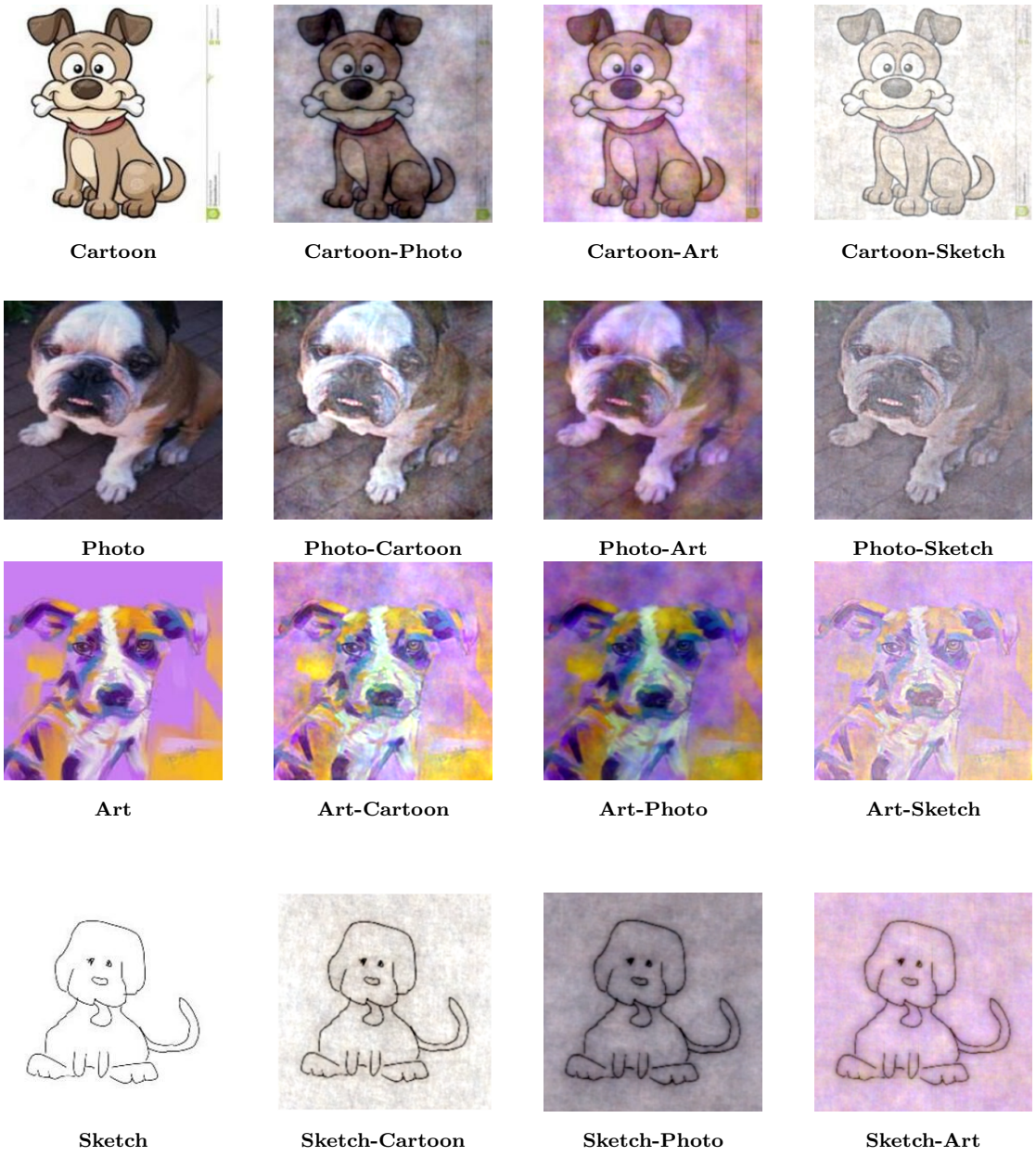

Figure 4.4 Augmenting images from the the PACS dataset during the training for the dog class.

In the context of APA, where augmented images are synthetically generated via frequency-space manipulation, improper scheduling may lead to adverse outcomes either by overwhelming the network with overly synthetic examples or by not exposing it to enough domain variability. Therefore, identifying the right balance between original and augmented data over time becomes essential for achieving optimal generalization performance.

The ablation study was carried out using the ResNet-50 architecture on the PACS dataset, employing the leave-one-domain-out evaluation protocol described earlier. I systematically varied both the step interval and the augmentation application ratio.

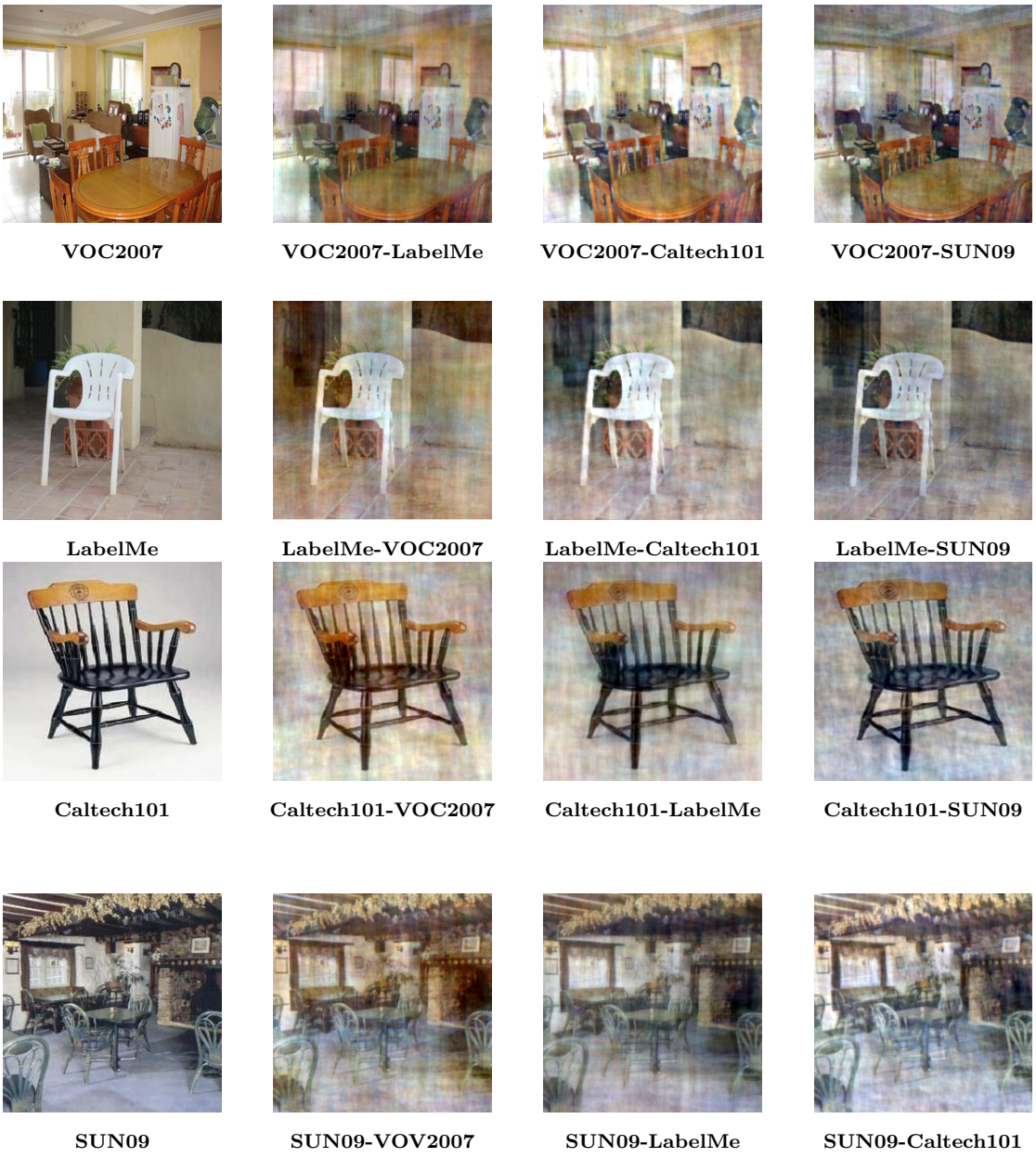

Figure 4.5 Augmenting images from the VLCS dataset during the training for the chair class.

The combinations I tested included step intervals of {50, 200, 500} and ratios of {0.1, 0.3, 0.5, 0.9}. Each combination was evaluated independently, and the results were plotted in Fig. 4.6, which visualizes the average classification accuracy across all target domains.

The results showed that APA's effectiveness was highly sensitive to the scheduling of augmentations. Specifically, when the step interval was set to 50 and the application ratio was as low as 0.1, the model exhibited significantly lower accuracy. This can be attributed to the infrequent occurrence and short duration of the augmentation phases, which caused the model to learn almost entirely from unaltered

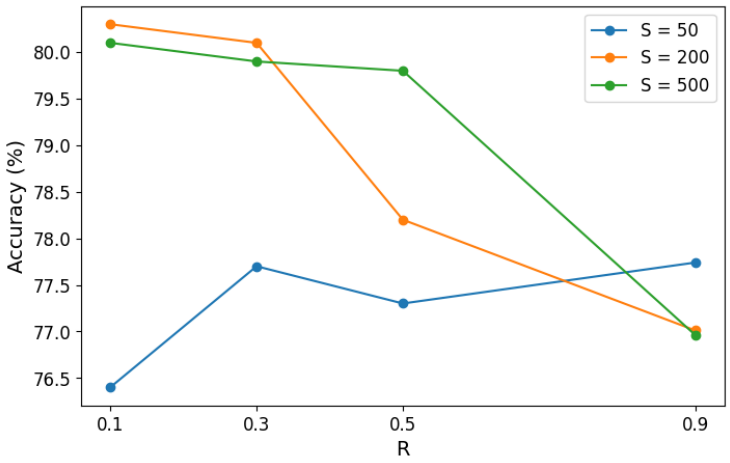

Figure 4.6 Accuracy for different values of S and R. S denotes the augmentation step interval, and R represents the augmentation application ratio on APA method using ResNet-50 as the backbone and PACS as the dataset.

source domain data. Consequently, the network failed to develop domain-invariant representations, leading to poor generalization.

On the other extreme, when the step interval was extended to 500 steps, the model experienced a lower frequency of augmentation phases but with longer durations during each phase. While this allowed deeper exposure to augmented samples during the active augmentation window, the long intervals between augmented batches caused the model to alternate sharply between overfitting and overgeneralization. The inconsistent learning pattern resulted in suboptimal convergence behavior.

The best performance was observed with a step interval of 200 and an application ratio of 0.1, which created a stable training rhythm where augmented and non-augmented samples complemented each other. These findings suggest that augmentation should be applied neither too rarely nor too frequently, and the transition between original and augmented data must be smooth and rhythmic to maintain learning stability.

in Fig. 4.7 you can see the same study on the DeiT-Small backbone and PACS dataset. Here I tried different range of values and got almost the same results, here I observed that I get the best performance using 200 as my interval steps but despite the previous ResNet-50 backbone, I got better results as I increased the R, it means that comparing to the ResNet-50, in a transformer backbone like DeiT-Small I need to feed the network with more portion of augmented images and it has more capacity to be fed with these synthetic images.

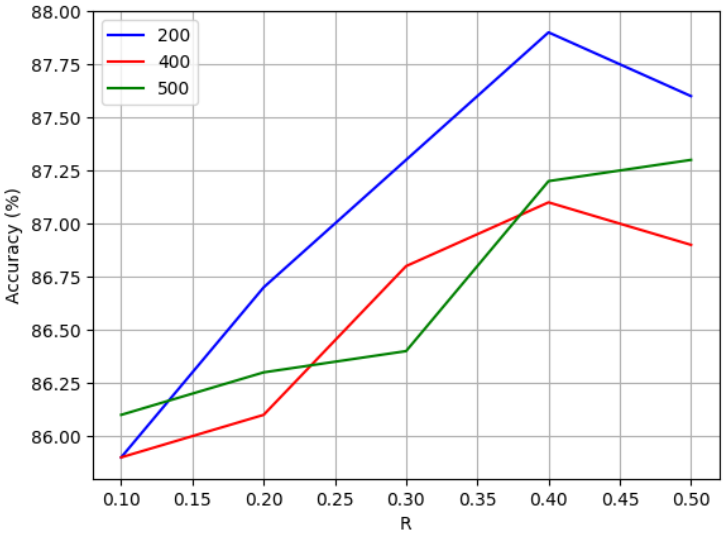

Figure 4.7 Accuracy for different values of S and R. S denotes the augmentation step interval, and R represents the augmentation application ratio on APA method using DeiT-Small as the backbone and PACS as the dataset.

4.1.6.2 Effect of Temporal Placement of Augmentation Steps

In a complementary set of experiments, I further investigated the impact of the temporal positioning of augmentation within each periodic training interval using ResNet-50 as the backbone and PACS and VLCS datasets. By default, APA applies augmentation during the final 20 steps of every 200-step interval. To explore whether this placement matters, I tested three scenarios where the same number of augmented steps (i.e., 20 steps) were positioned either at the beginning (steps 0 to 20), middle (steps 90 to 110), or end (steps 180 to 200) of each interval.

The results, reported in Table 4.12, showed a clear trend. The best performance was achieved when the augmentation was applied during the last segment of each training interval. When augmentation was applied at the beginning, the model was exposed to synthetic examples too early before learning meaningful representations from the source data. This caused the network to anchor its learning around less structured, potentially misleading inputs, thereby reducing its capacity to generalize. Augmenting in the middle of the interval led to marginally better performance, but still fell short of the default setting. These findings align with the curriculum learning principle, where the model benefits from a gradual increase in task difficulty: starting with real, clean data and then progressing toward more diverse and challenging examples.

I followed this temporal setting for the other experiments on the different backbones and datasets as well. From these ablation studies, several practical guidelines emerge

Table 4.12 Classification accuracy (%) on PACS and VLCS using ResNet-50 in different augmentation scenarios during total 200 steps in each training period.

Augmentation step	PACS	VLCS
step 0	83.3	79.4
step 90	83.8	79.7
step 180 (default)	**85.7**	**80.3**

for effectively deploying APA in training pipelines:

- **Moderate and rhythmic augmentation improves generalization**: Frequent but short bursts of augmentation are ineffective, and overly long augmentation windows can lead to overgeneralization. A periodic scheduling mechanism with balanced duration and frequency yields optimal results.
- **Delayed exposure to augmentation aids learning**: Augmentation applied at later training stages within each interval allows the model to first build structural priors from clean images, making subsequent exposure to frequency-altered data more beneficial.
- **APA is sensitive to timing**: Unlike random data augmentation strategies that apply uniformly throughout training, APA performs best when augmentation is introduced in a structured and temporally-aware manner.

5. CONCLUSION

In this thesis, I tackled the persistent and challenging problem of domain shift in the DG (Muandet et al., 2013). Unlike traditional DA techniques, DG assumes no access to target domain data during training, which necessitates the development of models that are inherently robust to distributional discrepancies across domains. To address this problem, I proposed a novel data augmentation strategy grounded in frequency-domain transformations using the FFT (Cooley & Tukey, 1965). This technique, APA, was designed to improve the generalization ability of deep learning models by synthesizing augmented views that reflect realistic inter-domain variations while preserving semantic fidelity.

The foundation of APA lies in transforming images from the spatial domain where conventional augmentations such as flipping, cropping, and color jittering operate into the frequency domain. Through the FFT, images are decomposed into two principal components: amplitude, which encodes structural and textural information, and phase, which captures spatial layout and semantic content. By perturbing the amplitude spectrum while preserving the phase, I ensured that the augmented images maintained the core identity of the original samples but presented them with modified global textures and structures. These modifications simulate the type of variations often encountered when models are deployed across unseen domains.

My method offers a lightweight, architecture-agnostic solution. The APA augmentation process injects diversity into the training data distribution without requiring any changes to the model or training pipeline. Moreover, by operating in the frequency domain, APA can explore a broader augmentation space that is difficult to capture using pixel-level operations. For instance, global texture shifts or periodic pattern alterations common in artistic styles, sensor artifacts, or environmental noise are better represented and manipulated in the frequency spectrum.

The motivation behind the proposed method is twofold. First, I sought to introduce a form of augmentation that better captures domain shifts encountered in real-world applications, where differences between domains are often subtle and embedded

in the texture, background statistics, or scene-level context. Second, I aimed to steer the learning process towards capturing domain-invariant representations, such as object shapes, contours, and relative spatial arrangements, by minimizing the influence of domain-specific cues.

To validate the effectiveness of APA, I performed extensive experiments on two widely adopted DG benchmarks: VLCS (Zhou et al., 2017) and PACS (Li et al., 2017). These datasets pose significant challenges due to their inter-domain variability, VLCS focuses on natural scene differences while PACS includes substantial stylistic transformations. Across both CNN-based (LeCun et al., 1998) and transformer-based backbones (Vaswani et al., 2017), my method demonstrated competitive results over baseline models, including state-of-the-art approaches such as MLDG (Li et al., 2018), DANN (Ganin et al., 2016), GMDG (Tan et al., 2024), and transformer-specific strategies like TFS-ViT (Noori et al., 2024) and SDViT (Sultana et al., 2022).

Moreover, through a detailed ablation study, I demonstrated the sensitivity of APA to its hyper-parameters and the importance of carefully scheduling augmentation during training. I showed that both the frequency and temporal placement of augmented steps play a critical role in maximizing generalization. These findings emphasize that, beyond the augmentation itself, its integration into the learning process must be strategically managed to yield the best outcomes.

Implications and Broader Impact

The implications of this work extend beyond the immediate results on benchmark datasets. Frequency-based augmentation techniques like APA offer a new perspective on how data can be manipulated to improve generalization. By shifting the focus from pixel-level variation to spectral-level diversity, my work encourages the broader machine learning community to consider frequency-aware training strategies, especially in domains where annotated data is limited or where domain shifts are substantial and difficult to model explicitly.

In practical terms, APA's architecture-independence makes it applicable to a wide range of models, including resource-constrained setups and industrial pipelines. Furthermore, because APA requires no target domain data or labels, it can be used in privacy-sensitive or deployment-critical applications where access to deployment domain data is not feasible.

Future Work

This thesis can be expanded in other ways. One natural extension is to apply APA to more complex and diverse datasets, such as those found in medical imaging, remote sensing, or low-light vision, where frequency-domain artifacts are more pronounced and domain shift is more severe. In such contexts, APA's ability to simulate domain variability without semantic distortion could prove especially valuable.

Another promising direction involves the development of adaptive frequency augmentation techniques. Rather than using fixed augmentation parameters, future work could explore dynamic adjustment of frequency transformations based on training progress, model confidence, or domain statistics. Such adaptive mechanisms could further enhance robustness while maintaining training efficiency.

Additionally, integrating APA with contrastive learning frameworks presents an exciting opportunity. By combining frequency-based augmentation with self-supervised or supervised contrastive objectives, models could be encouraged to learn disentangled and domain-invariant embeddings. This fusion has the potential to further improve cross-domain recognition and representation learning.

Finally, an intriguing research direction involves investigating the theoretical underpinnings of frequency-domain data augmentation. A deeper understanding of how spectral perturbations affect feature learning and representation geometry could help formalize best practices and inspire the creation of even more effective domain generalization techniques.

BIBLIOGRAPHY

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(12), e1006613.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, *19*, 137–144.

Bracewell, R. N. (1999). *The Fourier Transform and Its Applications* (3rd ed.). McGraw-Hill.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, (pp. 737–744).

Cai, M., Zhang, H., Huang, H., Geng, Q., Li, Y., & Huang, G. (2021). Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (pp. 13929–13938)., Montreal, Canada.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., & Park, S. (2021). Swad: Domain generalization by seeking flat minima. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 3019–3030)., Virtual. PMLR.

Choi, J., Seong, H. S., Park, S., & Heo, J.-P. (2023). Tcx: Texture and channel swappings for domain generalization. *Pattern Recognition Letters*, *175*, 74–80.

Cooley, J. W. & Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, *19*(90), 297–301.

Demirel, B., Aptoula, E., & Ozkan, H. (2023). Adrmx: Additive disentanglement of domain features with remix loss.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Everingham, M. & et al. (2010). Pascal voc challenge. *International Journal of Computer Vision*, *88*, 303–338.

Fei-Fei, L. & et al. (2003). Caltech-101: A caltech database for object category recognition.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Bach, F. & Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, (pp. 1180–1189)., Lille, France.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 1180–1189)., New York, NY, USA. JMLR.

Gonzalez, R. C. & Woods, R. E. (2007). *Digital Image Processing*. Prentice Hall.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Gulrajani, I. & Lopez-Paz, D. (2020). In search of lost domain generalization. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 1233–1244)., Virtual. PMLR.

Guo, H., Wang, Z., Zhang, Y., Sun, X., & Han, Y. (2023). Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 5509–5518)., New Orleans, LA, USA. IEEE.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1735–1742).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning (ICML)*, (pp. 2790–2799).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kostkova, J., Flusser, J., Lebl, M., & Pedone, M. (2020). Handling Gaussian blur without deconvolution. *Pattern Recognition*, *103*, 107264.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, (pp. 1097–1105).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, G., Jang, W., Kim, J. H., Jung, J., & Kim, S. (2024). Domain generalization using large pretrained models with mixture-of-adapters.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z., & Wu, Y. (2021). Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*.

Li, B., Shen, Y., Yang, J., Wang, Y., Ren, J., Che, T., Zhang, J., & Liu, Z. (2023). Sparse mixture-of-experts are domain generalizable learners. In *The Eleventh International Conference on Learning Representations*.

Li, D., Wang, Y., Gong, B., & Hospedales, T. M. (2019). Feature-critic networks for heterogeneous domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (pp. 1225–1234).

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (pp. 5542–5550).

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the International Conference on Machine Learning*, (pp. 2585–2594)., Stockholm, Sweden.

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 531–540)., Honolulu, HI, USA. IEEE.

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 10–18)., Atlanta, Georgia, USA. JMLR.

Nam, H., Lee, H., Park, J., Yoon, W., & Yoo, D. (2020). Reducing domain gap by reducing style bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2252–2261)., Seattle, WA, USA.

Noori, M., Cheraghalikhani, M., Bahri, A., Vargas Hakim, G. A., Osowiechi, D., Ayed, I. B., & Desrosiers, C. (2024). Tfs-vit: Token-level feature stylization for domain generalization. *Pattern Recognition*, *149*, 110213.

Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset Shift in Machine Learning*. MIT Press.

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European conference on computer vision*, (pp. 102–118). Springer.

Russell, B. C. & et al. (2008). Labelme: A database and web-based tool for image annotation.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 815–823).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, (pp. 618–626).

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Sultana, M., Naseer, M., Khan, M. H., Khan, S., & Khan, F. S. (2022). Self-distilled vision transformer for domain generalization. In *Asian Conference on Computer Vision (ACCV)*, volume 13842 of *Lecture Notes in Computer Science*, (pp. 286–302). Springer.

Sun, B. & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, (pp. 443–450)., Amsterdam, Netherlands. Springer.

Tan, Z., Yang, X., & Huang, K. (2024). Rethinking multi-domain generalization with a general learning objective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 23512–23522).

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun & L. Pratt (Eds.), *Learning to Learn* (pp. 181–209). Springer.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Learning Representations (ICLR)*.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, (pp. 5998–6008).

Wang, X., Zhang, J., Qi, L., & Shi, Y. (2025). Balanced direction from multifarious choices: Arithmetic meta-learning for domain generalization.

Wei, G., Lan, C., Zeng, W., Zhang, Z., & Chen, Z. (2021). Toalign: Task-oriented alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiao, J. & et al. (2010). Sun database: Large-scale scene recognition from abbey to zoo.

Yan, S., Song, H., Li, N., Zou, L., & Ren, L. (2019). Improve unsupervised domain adaptation with mixup training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (pp. 4161–4170)., Seoul, Korea.

Yang, Y. & Soatto, S. (2020). Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 4085–4095)., Virtual.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *IEEE International Conference on Computer Vision (ICCV)*, (pp. 558–567).

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, (pp. 6023–6032).

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., & Finn, C. (2020). Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, (pp. 1–12)., Vancouver, Canada. Curran Associates, Inc.

Zheng, K., Cao, Y., Zhu, K., Zhao, R., & Zha, Z. (2022). Famlp: A frequency-aware mlp-like architecture for domain generalization. *ArXiv*, *abs/2203.12893*.

Zhou, B., Song, S., Yu, L., Xiong, Y., & Wang, L. (2017). Vlcs dataset. Accessed: 2025-03-08.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *TPAMI*, *45*(4), 4396–4415.

Zhou, K., Yang, Y., Qiao, Y., & Loy, C. C. (2021). Domain generalization with mixstyle. In *International Conference on Learning Representations (ICLR)*.

APPENDIX A

Publications

The following is my publication based on this thesis:

- **Sina Salehnia**, Oznur Tastan, Erchan Aptoula. *Frequency Domain Image Augmentation for Domain Generalized Image Classification.* Accepted at **SIU 2025**.
- **Sina Salehnia**, Oznur Tastan, Erchan Aptoula. *APA: Domain Generalization Using Frequency Based Augmentation.* Accepted at **MLSP 2025**.