Denoising Pretrained Black-box Models via Amplitude-Guided Phase Realignment

Anonymous authors
Paper under double-blind review

Abstract

Pre-trained models tend to inherit noisy label information from their training datasets, internalising it as biased knowledge. While learning with label noise has been explored, existing approaches rarely address the mitigation of biased knowledge embedded in pre-trained representations introduced by noisy labels. Moreover, existing denoising methods invariably rely on modifying training datasets or models to improve downstream task performance. However, we observe a growing trend in which both pre-trained models and their training datasets are scaling up significantly and becoming increasingly inaccessible, making modifications ever more infeasible. In this paper, we propose a black-box biased knowledge mitigation method called "Lorem", which leverages feature frequency amplitudes to guide phase correction on pre-trained representations, without access to training data or model parameters. We first present empirical evidence that, across different noise levels, the phase components of pre-trained representations are more sensitive to noisy labels than the amplitude components, while discriminative information for classification is primarily encoded in the amplitude. Moreover, we find that the impact of noisy labels on amplitude is global, leading to a gradual loss of discriminative information. Therefore, corrective strategies must be adaptive across the entire frequency spectrum rather than limited to the high-frequency components. Inspired by this observation, we design a method that leverages the amplitude residual to realign phase, thereby removing biased knowledge from pre-trained representations. Experiments on a variety of popular pre-trained vision and language models suggest that, even with a simple linear classifier, our method can enhance downstream performance across a range of in-domain and out-of-domain tasks.

1 Introduction

As public expectations for the performance of pre-trained models continue to rise, to meet these expectations, pre-trained models increasingly rely on web-crawled data and large-scale annotations, which makes these large-scale pre-training datasets inevitably affected by noisy labels (Piktus et al., 2023). A growing concern is that noisy labels affect the learned weights of pre-trained models, leading to biased knowledge being encoded in their representations. For example, when benign tumours are mislabeled as malignant during pre-training, the model may overemphasise irrelevant pixel-level artefacts. This biased knowledge affects the generalisation and transferability of pre-trained models in downstream tasks, and addressing the problem by simply fine-tuning the representations with downstream data is often suboptimal. Since noisy labels alter the feature space, these distortions constrain the optimisation capacity of fine-tuning, especially when the downstream dataset is small or domain-shifted.

Existing research in learning with noisy labels has mainly focused on the training phase, with mainstream methods generally categorised into noise modelling (Van Rooyen et al., 2015; Han et al., 2018a; Van Rooyen & Williamson, 2018; Yao et al., 2019) and model-based techniques (Reed et al., 2014; Liu & Tao, 2015; Thulasidasan et al., 2019; Lyu & Tsang, 2019; Yu et al., 2019; Han et al., 2020). These methods assume access to the model's training data, architecture, and parameters, allowing them to mitigate the effects of label noise to some extent. However, such strategies are difficult to apply to improving the performance of large-scale pre-trained models on downstream tasks. Due to their size, re-training pre-trained models is often computationally infeasible—even when noisy labels are identified. Moreover, both the datasets and models are not always publicly available. Therefore, it remains an open problem to design effective strategies for mitigating the influence of noisy labels in pre-trained models in a black-box setting, where neither the original data nor model parameters are accessible.

In this paper, we explore a frequency-based approach to mitigating biased knowledge in pre-trained representations, leveraging the Fourier transform—a global transformation that decomposes representations into frequency components. This decomposition allows us to disentangle amplitude (global energy distribution) and phase (structural layout), enabling us to identify and correct frequency-specific distortions introduced by noisy supervision, thereby mitigating biased knowledge. We begin by presenting an empirical analysis, where amplitude and phase signals obtained under different levels of noisy supervision are used as predictive features, and the performance of classifiers built on these features is compared across downstream tasks. Our analysis shows that both amplitude and phase are affected as the noise level increases. However, classifiers built on amplitude signals consistently outperform those using phase, suggesting that phase should be the primary target of correction, as restoring distorted phase components is key to recovering the semantic integrity of pre-trained representations and improving their quality for downstream tasks. To further investigate, we examined how noisy labels affect amplitude and found that as the noise level increases, the magnitude, structural patterns, and distribution of amplitude have significant global changes, rather than being limited to the high-frequency components emphasised in prior work (Chen et al., 2021). These fluctuations imply that amplitude encodes label-related global patterns that are progressively distorted by noisy supervision. Thus, while amplitude itself is affected, its variations provide a reliable signal of where and how noise alters the feature space, making it a natural guide for phase realignment, where amplitude residuals are used to realign phase components, thereby mitigating biased knowledge in pre-trained representations.

Motivated by these findings, we propose "Lorem", a lightweight black-box framework designed to improve the generalisation ability of pre-trained representations. We first transform pre-trained features into a label-guided feature space, while preserving their original amplitude information. We then compute the amplitude residuals—defined as the difference between the amplitudes in the original feature space and those in the new feature space—and use them to guide the correction of the phase components. The aligned phase, along with the new amplitude, is combined via the inverse Fourier transform to generate enhanced representations, aiming to remove biased knowledge and support more reliable transfer to downstream tasks.

The main contributions of this paper are as follows:

- We provide an empirical analysis of amplitude and phase components under varying levels of label noise, demonstrating their differences in sensitivity to noisy labels and their relevance to discriminative information. Furthermore, subsequent experiments reveal that label noise globally changes both the structural patterns and the magnitude of amplitude. This highlights the potential of using amplitude to guide the correction of phase for mitigating biased knowledge in pre-trained representations.
- We propose a lightweight black-box method, named "Lorem", which uses amplitude residuals to guide the adjustment of phase. The enhanced features are then reconstructed via the inverse Fourier transform, allowing for improved generalisation and transferability of pre-trained models with minimal computational cost.
- We conduct extensive experiments on a variety of popular pre-trained vision and language models, evaluating both in-domain (ID) and out-of-domain (OOD) downstream tasks. Our method is compared against strong baselines and widely-used fine-tuning techniques, consistently demonstrating superior performance.

2 Problem Formulation

We assume the pre-training dataset is defined as $D = \{x_i, y_i\}_{i=1}^{N}$, where it consists of inputs $x_i \sim \mathcal{X}$ and supervision $y_i \sim \mathcal{Y}$, and the dataset size is N. The inputs and corresponding supervision can take any form. For example, in an image classification task, the inputs are images and the supervision are class labels (Russakovsky et al., 2015; He et al., 2016); in a sentence-pairing task, both the inputs and outputs are sentences (Cer et al., 2017; Wang et al., 2018). We further assume that the pre-training dataset contains noisy labels, such as data samples annotated with incorrect labels, which is common in large-scale pre-training datasets collected using web crawlers and similar techniques.

Thus, we define the noisy dataset as $\tilde{D} = \{x_i, \tilde{y}_i\}_{i=1}^{N}$, which we assume is inaccessible. The pre-trained model can be abstractly viewed as a combination of a feature extractor and a projection head. The pre-trained representations $\mathbf{F}$ are extracted from the feature extractor f_ϕ, and then mapped into the target domain using the projection head g_θ. Similarly, the parameters of the feature extractor f_ϕ are also assumed to be inaccessible. Since the training dataset contains noise, the extracted pre-trained representations may not accurately reflect the semantic content of the samples. Therefore, our goal is to repair the pre-trained representations to enhance the performance of pre-trained models on downstream tasks, without requiring re-training. To achieve this, we assume access to a downstream dataset $\mathcal{D}' = \{x_i, y_i\}_{i=1}^{K}$, which is used for optimization. Unless otherwise stated, we consider $\mathcal{D}'$ to be clean (i.e., without additional label noise).

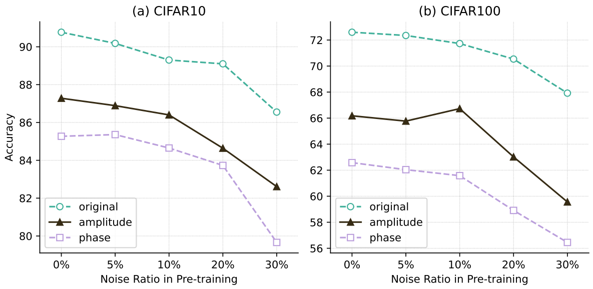

Figure 1: Classification accuracy on CIFAR-10 (a) and CIFAR-100 (b) using original pre-trained representations, their amplitude, and their phase under varying noise ratios in pre-training.

3 Analysis of frequency component robustness across noise levels

Given current amplitude–phase recombination approaches for learning with noisy labels (Huang et al., 2023; Chen et al., 2021), we are interested in investigating the downstream performance of the frequency components—amplitude and phase—of pre-trained representations obtained from models under different levels () of noisy supervision. In this analysis, we use five noisy pre-trained ResNet-50 (He et al., 2016) models from Chen et al. (2024), each trained with different levels of label noise, each trained with a different label noise ratio (0%, 5%, 10%, 20%, and 30%, where 0% denotes the clean dataset) using CNN backbones, to extract features from two image datasets: CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). We then apply the Fourier transform to obtain their amplitude and phase components, and compare their performance in linear probing.

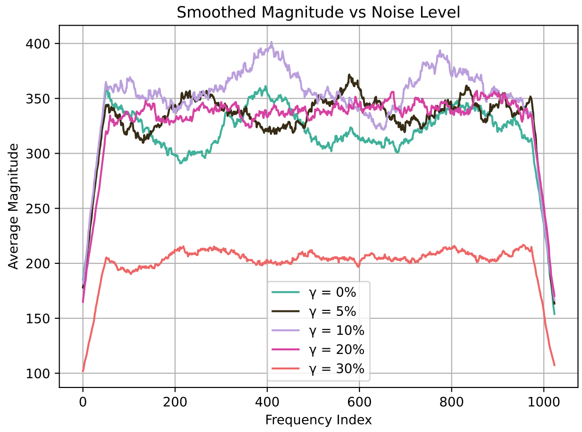

Figure 2: Smoothed amplitude magnitude across frequency indices under different noise ratios (γ). We divide the spectrum into three ranges based on the frequency index: low frequency (0–200), mid frequency (200–600), and high frequency (600–1000).

As shown in Figure 1, We have several observations:

(1) The classifiers using amplitude signals consistently outperform those using phase features in downstream classification tasks. This can be attributed to the pre-trained model's inherent tendency to encode label-related semantic information into the amplitude. The most visually obvious differences between categories—such as colour intensity, edge thickness, and texture density—are directly associated with the amplitude. During training, the model tends to prioritise memorising such easily accessible informa-

tion to converge more quickly, thereby embedding most of the discriminative information into the amplitude. Previous studies have also reported that CNNs favour amplitude over phase (Chen et al., 2021).

(2) Second, on CIFAR-100, amplitude-based classifiers using 0%–10% noisy pre-trained representations show minimal performance differences, and at the 10% level there is even a slight improvement. This performance trend may be attributed to noisy labels inducing the pre-trained model to focus more on local features. Under mild noise, the model may start associating a label with unique details of a specific image (e.g., reflections or the edge of a particular pixel), and memorising additional textures may accidentally increase discriminability, explaining why the 5% noisy model in Chen et al. (2024) slightly outperforms the clean model in downstream tasks.

(3) In contrast, phase-based classifiers are more sensitive to noisy labels, resulting in poorer classification performance. The amount of class-discriminative information contained in the phase is relatively limited. Although positional and structural information of the image subject is indeed encoded in the phase, such geometric structures do not always directly reveal class distinctions. For instance, in low-resolution datasets like CIFAR-10/100, the skeletal structures of cats and dogs are highly similar. On the other hand, in the phase domain, this local-focus bias causes small-scale shifts, which make the phase distribution more dispersed. This distorts the geometric information in the phase and further degrades classification.

Overall, from the classification results on the two image datasets, we find that the amplitude of pre-trained features encodes more label-related discriminative information, whereas the phase is more sensitive to noisy labels. They severely disrupt positional and structural information, leading to a more pronounced performance degradation for phase-based classifiers. To further investigate the impact of noisy labels on amplitude, we present the variation of the average amplitude magnitude of pre-trained representations from the CIFAR-10 dataset across different noise levels.

As shown in Figure 2, we observe that, compared to clean pre-trained representations, the first amplitude peak under 5% noisy supervision shifts from the low-frequency range (frequency index 0–200) to the mid-frequency range (frequency index 200–400). The second peak, which originally appears around frequency index 400, is also delayed and emerges near index 600, while the changes in the high-frequency range remain minor. For amplitudes obtained from 10% noisy supervision, both peaks are further shifted, with the first occurring around frequency index 400 and the second also around index 600. In addition, compared to the clean amplitude, the overall magnitude is noticeably higher. At extreme noise ratios of 20% and 30%, the structure becomes disrupted, with no distinct peaks. At the 30% noise level in particular, the overall magnitude level is even reduced, indicating severe loss of discriminative information. Overall, we argue that the impact of noisy labels does not always manifest first as a simple high-frequency peak (Huang et al., 2023), but rather as a global alteration of the amplitude magnitude structure. As the noise level increases, we can clearly observe the gradual shift of magnitude in the feature amplitude from low frequencies toward mid- and high-frequency ranges, accompanied by progressively weakened structural patterns. Therefore, our correction method does not need to target any specific frequency band; instead, by leveraging amplitude residuals, it can more effectively strengthen or suppress particular frequencies as needed.

4 Amplitude-Guided Phase Realignment

Motivation. Through experiments presented in Section 4, we observe that: (1) discriminative information is primarily encoded in the amplitude, while the phase is highly sensitive to noisy labels; and (2) noisy labels in the pre-training dataset reshape both the distribution and the structural patterns of amplitude in pre-trained features. Since the distribution patterns of amplitude correspond to specific spatial structures, these structures in turn determine the distribution of phase. For example, repetitive textures (e.g., brick walls, grids) produce periodic peaks in the frequency domain, and the arrangement of phases at these frequencies decides the alignment of textures. Variations in amplitude also reflect changes in texture repetitiveness; if the periodic peaks are disrupted, it often indicates phase misalignment (Yu et al., 2022). Based on these observations, amplitude variations can serve as an effective indicator for guiding phase correction. The adjusted phase better aligns with the semantic content of the samples, eliminates the biased knowledge embedded in pre-trained features, and enhances generalisation on downstream tasks.

Algorithm 1: Amplitude-Guided Phase Realignment Classifier (Lorem)

```
Input       : Batch features $\mathbf{X} \in \mathbb{R}^{B \times d}$
Parameters: MLP weights $(W_1, W_2)$ with BN+ReLU; classifier $W_c$; learnable template $\phi \in \mathbb{R}^{B \times d}$ (init
              $\pi$); step size $\varepsilon = 0.01$
Output      : Logits $\mathbf{M} \in \mathbb{R}^{B \times C}$
Semantic branch: $\mathbf{S} \leftarrow \mathrm{MLP}(\mathbf{X})$
Fourier transforms (last dim):
$\mathbf{X}' \leftarrow \mathcal{F}(\mathbf{X}), \quad \mathbf{S}' \leftarrow \mathcal{F}(\mathbf{S})$ $A_{\mathbf{X}} \leftarrow |\mathbf{X}'|, \quad \Phi_{\mathbf{X}} \leftarrow \angle(\mathbf{X}')$
$A_{\mathbf{S}} \leftarrow |\mathbf{S}'|, \quad \Phi_{\mathbf{S}} \leftarrow \angle(\mathbf{S}')$
Amplitude residual:
$r \leftarrow \tanh(A_{\mathbf{X}} - A_{\mathbf{S}})$
Phase update:
$\Delta\Phi \leftarrow r \odot \phi \quad \widehat{\Phi} \leftarrow \Phi_{\mathbf{S}} + \varepsilon\,\Delta\Phi$
Reconstruction (inverse Fourier transform):
$\hat{\mathbf{Z}} \leftarrow \Re\Big(\mathcal{F}^{-1}\big(A_{\mathbf{S}} \odot e^{i\widehat{\Phi}}\big)\Big)$
Classification:
$\mathbf{M} \leftarrow \hat{\mathbf{Z}} W_c^{\top}$
return $\mathbf{M}$
```

4.1 Method

Our method aims to leverage amplitude residuals in the frequency domain to guide phase correction, thereby improving the generalisation of pre-trained representations on downstream tasks. Specifically, given an input batch $\mathbf{X}$, we extract the original pre-trained representation $\mathbf{F}$. We then transfer this representation into a new semantic space through a two-layer multi-layer perceptron (MLP) h_ω, obtaining a new pre-trained representation $\mathbf{S}$. To enable comparison in the frequency domain, we apply the Fourier transform to both F and $\mathbf{S}$:

$$\mathbf{X}' = \mathcal{F}(\mathbf{F}) = A_{\mathbf{X}} \odot e^{i\Phi_{\mathbf{X}}}, \quad \mathbf{S}' = \mathcal{F}(\mathbf{S}) = A_{\mathbf{S}} \odot e^{i\Phi_{\mathbf{S}}}. \tag{1}$$

where $A_{\mathbf{X}}, A_{\mathbf{S}} \in \mathbb{R}^{B \times d}$ denote the amplitudes, and $\Phi_{\mathbf{X}}, \Phi_{\mathbf{S}} \in [-\pi, \pi]^{B \times d}$ denote the phases. Next, we compute the residual between the original and semantic amplitudes and regularise it with a $tanh(\cdot)$ function to ensure a stable training:

$$r = \tanh(A_{\mathbf{X}} - A_{\mathbf{S}}), \quad r \in (-1, 1)^{B \times d}. \tag{2}$$

Since the phase is an angular variable with values in $[-\pi, \pi]$, its update involves not only the magnitude of correction but also the sign (direction). However, the amplitude residual r only reflects the strength of the discrepancy between the two representations, lacking guidance on the direction of phase adjustment. To address this issue, we introduce a learnable phase template $\phi \in [-\pi, \pi]^{B \times d}$. By combining the residual r with ϕ, the model can adaptively learn which components require stronger correction and in which direction the correction should be applied. In other words, ϕ provides the phase correction with frequency selectivity, avoiding the under- or over-correction caused by uniform updates across all frequencies. Moreover, its learnability provides task adaptivity, enabling the model to automatically learn correction patterns that best fit a specific downstream task or data distribution, rather than relying solely on residual-driven updates. The phase correction term is given by the elementwise product of the amplitude residual r and ϕ, and can be represented as follows:

$$\Delta\Phi = r \odot \phi \tag{3}$$

The updated phase is then computed as:

$$\hat{\Phi} = \Phi_{\mathbf{S}} + \varepsilon \cdot \Delta\Phi, \quad \hat{\Phi} \in \mathbb{R}^{B \times d} \tag{4}$$

where $\varepsilon = 0.01$ is a global scaling factor that ensures gradual adjustments and stabilises training. Here, we use Φ_S instead of Φ_X because Φ_X comes directly from pre-training under noisy supervision, and therefore inherently contains biased knowledge. If used directly, it would easily bring noise into the correction process.

Finally, the updated phase is combined with the semantic amplitude, and the inverse Fourier transform is applied to reconstruct the corrected representation:

$$\hat{\mathbf{Z}} = \Re\Big(\mathcal{F}^{-1}\Big(A_{\mathbf{S}} \odot e^{i\hat{\Phi}}\Big)\Big), \quad \hat{\mathbf{Z}} \in \mathbb{R}^{B \times d}. \tag{5}$$

The operator $\Re(\cdot)$ denotes taking the real part to ensure that the reconstructed representation remains in the real space $\mathbb{R}^d$. The corrected representation $\hat{\mathbf{Z}}$ is then fed into a linear classifier W_c, and the entire model is optimised under the supervision of downstream labels.

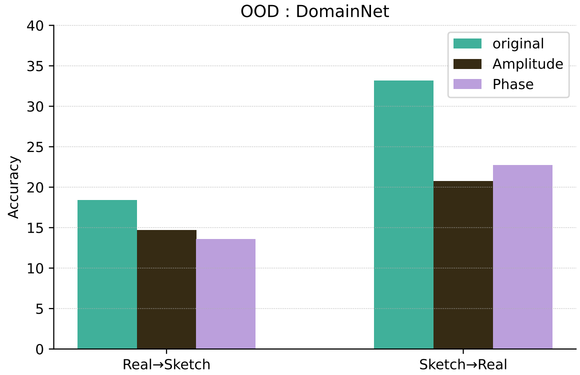

Figure 3: Classification accuracy on DomainNet using original pre-trained representations, their amplitude, and their phase when training on the "real" subset and testing on the "sketch" subset, and vice versa.

4.2 Discussion

One may argue that Lorem may fail if class-related texture patterns are not captured in the amplitude. To further investigate the effectiveness of our proposed method, we follow the experimental setup of Section 3 and design two types of OOD tasks: (1) Using the "real" subset of DomainNet as the training set and the "sketch" subset as the testing set, and vice versa; (2) Using the original CIFAR-10/100 training sets while applying four types of blur perturbations—Gaussian blur, motion blur, glass blur, and defocus blur—to the corresponding test sets.

As shown in Figure 3, when the "real" subset of DomainNet is used for training, the amplitude-based classifier slightly outperforms the phase-based classifier. However, when the roles are reversed, the phase-based classifier performs better than the amplitude-based one. This difference stems from the fact that the decision boundaries learned from the "real" subset rely heavily on texture information encoded in the amplitude. Although such information is removed in the "sketch" subset through decolorisation and texture removal, the overall shapes remain, thereby preserving transferability. When the training set becomes the "sketch" subset, which only contains positional and structural information, the amplitude-based classifier merely learns simple colour and texture patterns, causing it struggle when applied to the more complex "real" subset. Figure 4 shows the OOD results under the four blur types. The amplitude-based classifier outperforms the phase-based classifier in all but the glass blur setting, which swaps local pixel pairs, reassigning class-specific texture patterns, but leaves object edges and geometry unaffected; as a result, the phase retains some discriminative information. The other three types of blur mainly reduce fine-grained texture details while preserving the basic texture type, enabling amplitude-based classifier to remain effective.

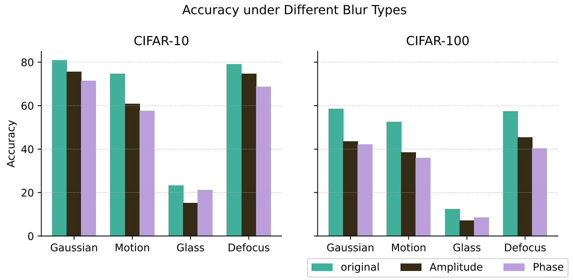

Figure 4: Classification accuracy on CIFAR-10/100 using original pre-trained representations, their amplitude, and their phase, when four types of blur perturbations are applied to the test sets.

Table 1: Performance comparison on ID tasks. Evaluation metrics: (1) "Acc" denotes accuracy; (2) "MPC" denotes mean per class.

Model	Tuning	**CIFAR-10**	**CIFAR-100**	**Caltech101**	**Food101**	**EuroSAT**	**RESISC45**	**StanfordCars**	**Avg**
		Acc	Acc	MPC	Acc	Acc	Acc	Acc	
Resnet-50	LP	81.40	59.60	88.00	49.95	93.27	79.37	32.10	69.10
	MLP	82.36	59.07	87.08	55.96	94.71	84.28	46.65	72.87
	NMTune	81.37	55.75	84.98	52.45	94.74	82.56	42.78	70.66
	ours	**82.46**	**59.95**	**88.23**	**56.75**	**95.03**	**85.18**	**48.46**	**73.72**
EfficientNet-B3	LP	89.88	69.55	89.68	67.52	95.50	86.77	48.02	78.13
	MLP	92.14	72.53	92.25	**72.18**	96.53	88.70	57.85	81.74
	NMTune	91.97	70.22	91.71	69.57	96.27	87.43	58.98	80.88
	ours	**92.19**	**72.85**	**92.63**	72.11	**96.64**	**89.33**	**61.38**	**82.45**
Swin-L	LP	95.68	82.31	93.83	85.05	96.53	87.79	50.78	84.57
	MLP	96.42	84.28	95.72	87.23	97.23	90.36	64.19	87.92
	NMTune	96.39	82.67	95.42	85.69	97.16	89.38	64.27	87.28
	ours	**96.48**	**84.45**	**96.18**	**87.26**	**97.33**	**90.67**	**69.72**	**88.87**
Resnet-50-5%	LP	88.68	68.76	84.42	57.41	94.37	83.84	34.27	73.11
	MLP	91.35	72.17	90.29	**63.19**	95.96	86.65	49.11	78.39
	NMTune	91.28	69.87	87.75	60.09	95.88	85.68	49.28	77.12
	ours	**91.42**	**72.54**	**91.00**	63.14	**96.12**	**87.69**	**51.90**	**79.12**
Resnet-50-10%	LP	88.28	69.12	83.38	56.14	94.59	84.02	34.34	72.84
	MLP	91.14	72.16	87.60	**61.98**	95.61	87.04	49.09	77.80
	NMTune	90.96	69.70	85.51	59.17	**95.88**	85.89	49.55	76.67
	ours	**91.30**	**72.37**	**87.99**	61.83	95.68	**88.01**	**52.39**	**78.51**

In summary, our proposed method is able to achieve superior performance in most downstream tasks because the amplitude captures class-related texture patterns with relatively high generalisability. As long as such patterns are not completely destroyed in the target domain, the model can still extract relevant class-discriminative information.

5 Experiments

Pre-trained models. For vision pre-trained models, we select the fully supervised ResNet-50 (He et al., 2016) and Swin-L (Liu et al., 2021b), as well as the semi-supervised EfficientNet-B3 (Tan & Le, 2019). In addition, we include two noisy ResNet-50 pre-trained models trained on ImageNet-1K (IN-1K) (Russakovsky et al., 2015) with 5% and 10% label noise, provided by Chen et al. (2024). In their setup, noisy supervision is simulated by uniformly flipping ground-truth class labels into other classes. For text pre-trained models, we select BERT-Large (Devlin et al., 2019), RoBERTa-Large (Liu et al., 2019), and GPT-2 (Radford et al., 2019). Both BERT-Large and RoBERTa-Large are Transformer encoders, while GPT-2 is a Transformer decoder pre-trained on WebText using large-scale autoregressive language modelling.

Datasets. We validate our model on seven in-domain (ID) vision tasks and two out-of-domain (OOD) vision tasks. For the ID tasks, we use seven downstream datasets, including CIFAR-10/100 (Krizhevsky et al., 2009), Caltech101 (Fei-Fei et al., 2004), Food101 (Bossard et al., 2014), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), and StanfordCars (Krause et al., 2013). For the OOD tasks, we use the "real" subset of DomainNet (Peng et al., 2019) as the training set and the "sketch" subset as the testing set, and vice versa. For text tasks, we validate our model on GLUE (Wang et al., 2018) and GLUE-X (Yang et al., 2022), for both ID and OOD evaluation.

Baselines. We compare our method with three related approaches: (1) Linear Probing: a single fully connected layer (Linear-Probe (FC)) and a two-layer MLP classifier (Linear-Probe (MLP)); (2) NMTune (Chen et al., 2024): a black-box method that applies singular value decomposition (SVD) to the original pre-trained features and introduces three regularisation strategies to adjust the Singular Value Entropy (SVE) and Largest Singular Value Ratio (LSVR), thereby transferring pre-trained representations into the new feature space and improving the generalisation of pre-trained models on downstream tasks.

Table 2: Performance comparison on OOD tasks. Evaluation metric: (1) “Acc” denotes accuracy.

Model	Tuning	DomainNet Real	DomainNet Sketch	Avg
		Acc	Acc	
Resnet-50	LP	16.67	31.08	23.88
	MLP	19.69	32.03	25.86
	NMTune	17.61	28.00	22.81
	ours	**20.02**	**32.19**	**26.10**
EfficientNet-B3	LP	23.57	37.40	30.49
	MLP	**25.46**	**40.54**	**33.00**
	NMTune	23.03	35.10	29.07
	ours	**25.46**	37.70	31.58
Swin-L	LP	38.67	**60.50**	49.59
	MLP	**40.38**	59.65	**50.02**
	NMTune	37.48	58.64	48.06
	ours	39.65	60.34	50.00
Resnet-50-5%	LP	18.02	33.02	25.52
	MLP	19.91	**36.68**	**28.30**
	NMTune	18.20	34.02	26.11
	ours	**19.94**	36.22	28.08
Resnet-50-10%	LP	17.64	32.33	24.99
	MLP	19.90	**36.05**	**27.98**
	NMTune	17.87	33.30	25.59
	ours	**19.92**	35.40	27.66

Table 3: Performance comparison on GLUE tasks. Evaluation metrics: (1) “Acc” denotes accuracy; (2) “MCC” denotes matthews correlation; (3) “PCC” denotes pearson correlation.

Model	Tuning	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST	STS	Avg
		MCC	Acc	Acc	Acc	Acc	Acc	Acc	PCC	
BERT-L	LP	38.97	**32.50**	70.06	66.59	77.07	**56.75**	83.53	**76.27**	62.72
	MLP	41.90	32.37	67.93	**69.01**	**84.18**	55.67	85.64	74.57	63.91
	NMTune	40.11	31.55	69.10	68.94	83.80	53.93	86.31	76.16	63.75
	ours	**42.53**	**32.50**	**70.47**	68.12	83.82	55.67	**86.93**	75.43	**64.43**
RoBERTa-L	LP	46.84	31.35	**72.07**	70.32	78.37	58.05	**87.84**	66.78	63.95
	MLP	45.46	35.85	71.23	70.64	84.34	54.37	85.07	68.96	64.49
	NMTune	39.15	35.45	69.26	72.05	84.48	55.67	86.95	68.07	63.88
	ours	**48.19**	**36.32**	71.32	**72.66**	**84.64**	**59.13**	86.63	**70.25**	**66.14**
GPT-2	LP	17.19	37.55	70.41	**67.24**	74.30	50.04	**83.30**	44.99	55.63
	MLP	20.15	37.06	69.63	66.68	**82.43**	51.26	76.17	58.43	57.73
	NMTune	16.31	38.49	66.72	63.03	80.91	50.54	58.06	58.03	54.01
	ours	**22.22**	**40.01**	**71.59**	66.69	80.70	**52.42**	81.36	**60.50**	**59.44**

Experimental Setting. We train each downstream classifier for 30 epochs using the Adam optimizer. The learning rate is set to 0.001 for the other baselines and 0.0001 for our proposed method. For evaluation, we report the average performance results over five runs.

5.1 Results

Vision tasks. Tables 1 - 2 show the results comparing our models to other baselines. We use the same datasets and identical train–validation–test splits, reporting the average accuracy across five runs

Table 4: Performance comparison on GLUE-X tasks. Evaluation metrics: (1) “Acc” denotes accuracy; (2) “MCC” denotes matthews correlation.

Model	Tuning	GT	IMDB	MNLI-mis	SNLI	SICK	NewsQA	SciTail	HANs	Avg
		MCC	Acc	Acc	Acc	Acc	Acc	Acc	Acc	
BERT-L	LP	42.64	72.08	52.77	70.39	53.36	37.21	59.16	49.60	54.65
	MLP	43.87	72.12	58.76	**74.39**	72.57	37.27	60.78	48.69	58.56
	NMTune	42.23	72.63	58.77	72.99	75.92	37.41	59.88	48.59	58.55
	ours	**45.45**	**72.69**	**59.29**	74.38	**77.09**	**38.83**	**60.94**	**50.07**	**59.84**
RoBERTa-L	LP	46.41	**71.77**	74.09	70.86	43.98	39.17	55.47	**49.99**	56.47
	MLP	45.66	59.45	75.25	71.80	50.10	38.95	61.31	49.75	56.53
	NMTune	39.81	59.23	76.35	72.67	50.83	38.67	60.43	48.82	55.85
	ours	**47.54**	70.65	**77.40**	**74.30**	**57.59**	**39.71**	**62.61**	49.86	**59.96**
GPT-2	LP	41.02	50.94	58.43	58.50	62.45	34.84	55.75	51.98	51.74
	MLP	40.51	**51.05**	**64.79**	63.32	57.49	37.72	55.34	52.92	52.89
	NMTune	42.48	50.63	57.88	65.13	57.30	37.20	41.51	52.76	50.61
	ours	**43.66**	51.04	63.76	**65.50**	**62.87**	**38.64**	**58.11**	**53.46**	**54.63**

on each dataset. Overall, our method consistently achieves higher accuracy in in-domain (ID) vision tasks, improving the quality of noisy pre-trained features. When using a single-layer linear network as the classifier, our approach significantly enhances the generalisation ability of pre-trained models on downstream tasks compared to standard linear probing. In out-of-domain (OOD) vision tasks, the performance of our method is mixed. Nevertheless, the gap between our method and the best-performing approach is small, and it still outperforms other baselines. The reason for this discrepancy has already been analysed in the discussion section (Section 4.2).

NLP tasks. Tables 3 - 4 present the comparison of our method against other baselines on GLUE and GLUE-X tasks. GLUE and GLUE-X can be regarded as the ID and OOD tasks in NLP, respectively. Overall, our proposed method achieves the best average performance across both types of tasks, demonstrating its ability to adapt to diverse downstream tasks and dataset distributions, thereby improving the generalisation of pre-trained models. Another finding is that the MLP classifier also performs strongly in both ID and OOD tasks, ranking just below our method. In contrast, simple linear probing, while occasionally strong on specific tasks, shows unstable performance overall, consistently ranking last in average results. This conclusion differs from that of Chen’s work, which we believe may be due to differences in learning rate selection.

6 Related works

Noisy Supervision. Mainstream approaches can be broadly categorised into the following directions: (1) Noise Modelling (Van Rooyen et al., 2015; Han et al., 2018a; Van Rooyen & Williamson, 2018; Yao et al., 2019): Directly modelling the label noise generation process to correct the loss during feature learning; (2) Robust Loss Functions (Liu & Tao, 2015; Zhang & Sabuncu, 2018; Thulasidasan et al., 2019; Charoenphakdee et al., 2019; Lyu & Tsang, 2019; Menon et al., 2020): Modifying the optimisation objective to enable the model to learn features even under noisy labels; (3) Regularisation (Reed et al., 2014; Azadi et al., 2015; Zhang et al., 2018; Han et al., 2020): Introducing prior constraints so that the features are less sensitive to noisy labels; and (4) Multi-model Learning (Veit et al., 2017; Li et al., 2017; Han et al., 2018b; Yu et al., 2019): Using multiple models to supervise each other, thereby reducing the risk of a single model being misled by noise. In contrast to these methods, our work focuses on mitigating the influence of biased knowledge in pre-trained representations caused by noisy labels. The most relevant to our work is Noisy Model Tuning (NMTune) proposed by Chen et al. (2024), a black-box fine-tuning method that employs multiple regularisation strategies. In Chen’s work, the authors compare Singular Value Entropy (SVE) and Largest Singular Value Ratio (LSVR) computed from pre-trained features obtained under different levels of noisy supervision, analysing how noise affects representation learning and generalisation. As the noise ratio increases, LSVR rises and SVE becomes excessively high, indicating that the pre-trained features become more constrained to specific directions and less diverse on new data distributions—ultimately reducing transferability. Based on these observations, Chen proposes adjusting the pre-trained feature space to reduce the effect of noise and improve generalisation. We note that there has been relatively little work on understanding the influence of biased knowledge in pre-trained features. Our work approaches this problem from a frequency-domain perspective, examining how different frequency components change as noise levels increase. In contrast to Chen's findings, our conclusion is that higher noise levels cause pre-trained features to learn high-frequency details and unique texture patterns, rather than concentrating on a few specific dominant directions. Thus, our work offers a distinct interpretation of the problem from the frequency-domain viewpoint, providing new insights into this research area.

Applications of Fourier Transform. Frequency-domain analysis helps neural networks identify and preserve key features more effectively. Several studies have provided new insights into explaining the behaviour of neural networks from a frequency-domain perspective (Wang et al., 2020; Guo et al., 2020; Chen et al., 2021; Liu et al., 2021a; Yu et al., 2022; Zhou et al., 2024). In the visual domain, some research has found that high-frequency components play an important role in improving the accuracy of CNNs (Wang et al., 2020; Chen et al., 2021). In Chen et al. (2021), the authors observed that CNNs tend to converge to local optima closely related to the high-frequency components of training images. However, the high-frequency components are susceptible to noise and perturbations. Inspired by the phenomenon in the human visual system—where robust recognition relies more heavily on phase information—the authors proposed a novel data augmentation method that recombines the phase spectrum of the current image with the magnitude spectrum of a distractor image to generate new training samples. This approach encourages CNNs to focus more on structural information (derived from phase) and become more robust to variations in amplitude (such as noise, brightness, and colour distortions). In Huang et al. (2023), the authors investigated the differential impacts of phase and amplitude on CNN robustness, proposing a method to decouple phase and amplitude in certain layers via the Discrete Fourier Transform (DFT) during training, and applying distinct early-stopping strategies to each component, thereby enhancing the network's robustness. In natural language processing (NLP), research has shown that pre-trained large language models implicitly utilise Fourier features when performing arithmetic tasks (Zhou et al., 2024). MLP and attention layers leverage low-frequency and high-frequency Fourier components, respectively, to accomplish tasks, and different pre-training strategies directly influence the model's effective utilisation of these Fourier features. Our work also investigates the impact of noisy labels on pre-trained features from a frequency-domain perspective. The key distinction from prior work is that we focus on leveraging the observed characteristics to improve generalisation in downstream tasks, rather than modifying the pre-trained model itself or improving its training process.

7 Conclusion

In this paper, we propose a novel black-box method to mitigate the impact of noisy labels on the downstream performance of pre-trained models. The method leverages amplitude residuals to realign the original phase of pre-trained representations, thereby mitigating biased knowledge and improving generalisation. By analysing the performance of amplitude and phase extracted from pre-trained representations trained under varying levels of noisy supervision, as well as the changes in amplitude distribution induced by noisy labels, we gain deeper insights into how noisy labels distort phase, drive models to overfit on irrelevant texture patterns, and consequently preserve biased knowledge in pre-trained representations. Our algorithm employs amplitude residuals as guidance for phase correction, making the representations more robust and generalisable. Experimental results demonstrate that our method outperforms state-of-the-art baseline methods and widely used fine-tuning approaches. We conduct experiments on a variety of vision and language pre-trained models, and our method achieves competitive results on both in-domain (ID) and out-of-domain (OOD) tasks. It is important to note that our approach operates in the frequency domain of pre-trained representations. Whether it can be extended to large language models and other foundation models remains an open challenge, which warrants further exploration in future research.

References

Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep cnns with noisy labels. *arXiv preprint arXiv:1511.07069*, 2015.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pp. 961–970. PMLR, 2019.

Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 458–467, 2021.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Uncertainty in Artificial Intelligence*, pp. 1127–1137. PMLR, 2020.

Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *Advances in neural information processing systems*, 31, 2018a.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018b.

Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Huaxi Huang, Hui Kang, Sheng Liu, Olivier Salvado, Thierry Rakotoarivelo, Dadong Wang, and Tongliang Liu. Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16719–16730, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1910–1918, 2017.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021a.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International conference on learning representations*, 2020.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. The roots search tool: Data transparency for llms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 304–314, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*, 2019.

Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 839–847, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8684–8694, 2020.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*, 2022.

Jiangchao Yao, Hao Wu, Ya Zhang, Ivor W Tsang, and Jun Sun. Safeguarded dynamic label regression for noisy supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9103–9110, 2019.

Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European conference on computer vision*, pp. 181–198. Springer, 2022.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pp. 7164–7173. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use fourier features to compute addition. *Advances in Neural Information Processing Systems*, 37:25120–25151, 2024.