Lightweight Frequency Masker for Cross-Domain Few-Shot Semantic Segmentation

Jintao Tong   **Yixiong Zou**[*]   **Yuhua Li**   **Ruixuan Li**
School of Computer Science and Technology, Huazhong University of Science and Technology
{jintaotong, yixiongz, idcliyuhua, rxli}@hust.edu.cn

Abstract

Cross-domain few-shot segmentation (CD-FSS) is proposed to first pre-train the model on a large-scale source-domain dataset, and then transfer the model to data-scarce target-domain datasets for pixel-level segmentation. The significant domain gap between the source and target datasets leads to a sharp decline in the performance of existing few-shot segmentation (FSS) methods in cross-domain scenarios. In this work, we discover an intriguing phenomenon: simply filtering different frequency components for target domains can lead to a significant performance improvement, sometimes even as high as 14% mIoU. Then, we delve into this phenomenon for an interpretation, and find such improvements stem from the reduced inter-channel correlation in feature maps, which benefits CD-FSS with enhanced robustness against domain gaps and larger activated regions for segmentation. Based on this, we propose a lightweight frequency masker, which further reduces channel correlations by an Amplitude-Phase Masker (APM) module and an Adaptive Channel Phase Attention (ACPA) module. Notably, APM introduces only 0.01% additional parameters but improves the average performance by over 10%, and ACPA imports only 2.5% parameters but further improves the performance by over 1.5%, which significantly surpasses the state-of-the-art CD-FSS methods.

1   Introduction

Recent advancements in semantic segmentation have been driven by large-scale annotated datasets and developments in deep neural networks [8, 29, 50, 44]. Nevertheless, the requirement for extensive labeled data remains a significant challenge, particularly for dense prediction tasks like semantic segmentation. Hence, few-shot semantic segmentation (FSS) [35, 11, 49] has been proposed to meet this challenge, aiming to produce predictions for the unseen categories with only limited annotated data. However, these FSS methods perform poorly when confronted with domain shifts, particularly when there is a significant gap between the novel class (target domain) and the base class (source domain). This issue has spurred the development of the cross-domain few-shot semantic segmentation (CD-FSS) task [24]. Despite various efforts in CD-FSS, the outcomes remain sub-optimal.

To handle the domain shift problem, efforts have been made to study the generalization of neural networks. Recently, some works [42, 40, 7] have explored this from the perspective of the frequency domain, achieving theoretical breakthroughs. Compared to humans, neural networks exhibit heightened sensitivity to different frequency components. Additionally, amplitude and phase exhibit distinct properties and effects on neural network performance. Inspired by these works, we study the domain shift problem from the perspective of the frequency domain and discover an intriguing phenomenon shown in Figure 1: for a model already trained on the source domain, **simply filtering frequency components of images during testing can lead to significant performance improvements**, sometimes even as high as 14% mIoU.

---

[*]Corresponding author.   Code is available at https://github.com/TungChintao/APM.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).
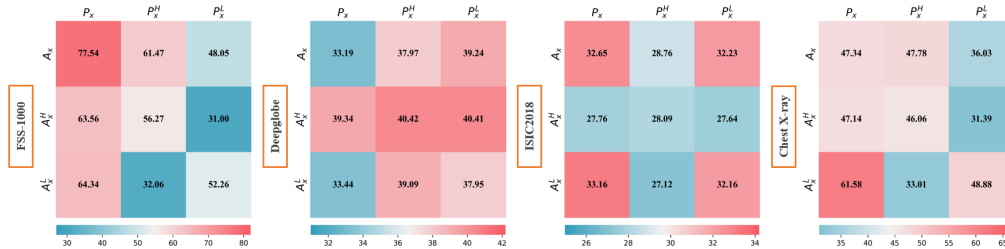
Figure 1: For a model already trained on the source domain, we simply filter out different frequency components and plot mIoU against the maintained ones of images. P denotes Phase, A denotes Amplitude, H denotes High Frequency, and L denotes Low Frequency. We can see the performance is significantly improved in most cases compared with the baseline (A_x, P_x), even as high as 14% on the Chest X-ray dataset (A_x^L, P_x). In this paper, we delve into this phenomenon for an interpretation, and propose a lightweight frequency masker for efficient cross-domain few-shot segmentation.

In this paper, we delve into this phenomenon for an interpretation. Through experiments and mathematical derivations, we find the filtering of the phase and amplitude effectively disentangles feature channels, which lowers the channel correlations and helps the model capture a larger range of semantic patterns. This benefits the model with improved robustness against large domain gaps, and helps to discover the whole object for segmentation.

Based on the above interpretations, we propose a lightweight frequency masker for the CD-FSS task. This masker does not need to be trained on the source domain, and can be directly inserted into intermediate feature maps during target-domain fine-tuning. It includes an Amplitude-Phase Masker (APM) module and an Adaptive Channel Phase Attention (ACPA) module. APM adaptively learns on target domains to filter out harmful amplitude and phase components at a finer granularity, which improves the effectiveness of channel disentanglement. ACPA learns the attention over channels through phase information. Notably, the APM module only introduces 0.01% additional parameters, but can effectively improve the mIoU by over 10% on average, and ACPA further improves the performance by 1.5% on average with only 2.5% additional parameters introduced.

In summary, our contributions can be listed as

- We find a phenomenon that simply filtering frequency components on target domains can significantly improve performance, with the highest improvement reaching nearly 14%.

- We delve into this phenomenon for an interpretation. We find the frequency filtering operation can effectively disentangle feature map channels, benefiting the model with improved robustness against large domain gaps and a larger range of discovered regions of interest.

- Based on our interpretations, we propose a lightweight frequency masker for the CD-FSS task, which significantly improves the mIoU by 11% on average with only 2.5% parameters introduced.

- Extensive experiments on four target datasets show that our work, through a simple and effective design, significantly outperforms the state-of-the-art CD-FSS method.

2 Interpreting Enhanced Performance from Frequency Filtering

In this section, we delve into why filtering certain frequency components can significantly improve CD-FSS performance in certain target domains for interpretation.

2.1 Preliminaries

Cross-domain few-shot semantic segmentation (CD-FSS) aims to generalize knowledge acquired from source domains with ample training labels to unseen target domains. Given a source domain $D_s = (\mathcal{X}_s, \mathcal{Y}_s)$ and a target domain $D_t = (\mathcal{X}_t, \mathcal{Y}_t)$, where $\mathcal{X}$ represents input data distribution and $\mathcal{Y}$ represents label space. The model will be trained on the training set from the D_s, then applied to perform segmentation on novel classes in the D_t. Notably, D_s and D_t exhibit distinct input data distribution, with their respective label spaces having no intersection, i.e., $\mathcal{X}_s \neq \mathcal{X}_t$, $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$.

In this work, we adopt the episodic training manner. Specifically, both the training set sampled from D_s and the testing set sampled from D_t are composed of several episodes, each episode is constructed

2

Table 1: Mutual information between feature channels for the best and the worst cases in Fig. 1. We find that mutual information (MI) consistently decreases when the performance is improved.

Dataset	FSS-1000			Deepglobe			ISIC			Chest X-ray		
	baseline	best	worst	baseline	best	worst	baseline	best	worst	baseline	best	worst
1-shot mIoU	77.54	64.34↓	48.054↓	33.19	40.42↑	33.44↑	32.65	33.16↑	27.124↓	47.34	61.58↑	31.394↓
support MI	1.3736	1.3791↑	1.8767↑	1.3679	1.35024↓	1.35584↓	1.3789	1.36974↓	1.3951↑	1.3952	1.39304↓	1.4315↑
query MI	1.3739	1.3805↑	1.8201↑	1.3667	1.34354↓	1.35984↓	1.3792	1.36944↓	1.3890↑	1.3921	1.38774↓	1.4368↑
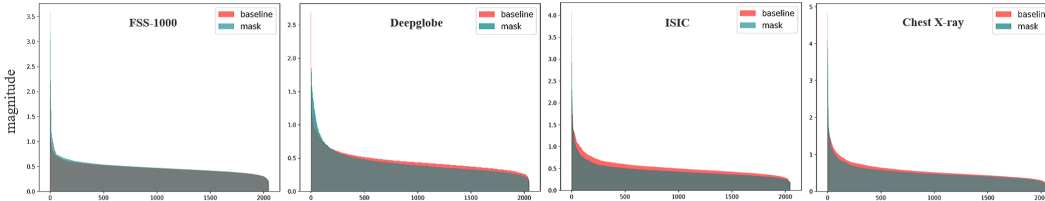
Figure 2: Mean Magnitude of Channels (MMC) for the best case in Fig. 1 on four target datasets. For domains with improved performance, their curves are lower than the baseline after masking.

of K support samples $S = \{I_s^i, M_s^i\}_{i=1}^K$ and a query $Q = \{I_q, M_q\}$ (I is the image and M is the label). Within each episode, the model is expected to use $\{I_s, M_s\}$ and I_q to predict the query label.

2.2 Enhanced Performance Stem from Reduced Inter-Channel Correlation

Existing research indicates inter-channel relationships are crucial for performance, as different feature channels can represent distinct features [3, 30]. Therefore, we study the change of channel correlations brought by the frequency filtering. Specifically, we measured the mean mutual information (MI) [4] between channels from the last layer of the backbone network. The measured cases include the combination of phase and amplitude for the highest and lowest performance in Fig. 1. We report the 1-shot mIoU and MI in Tab. 1, where we can see that for frequency combinations with improved mIoU, their MI consistently decreases, indicating reduced inter-channel correlation in the feature maps. Conversely, for frequency combinations with decreased mIoU, their MI consistently increases.

The higher the mutual information value, the larger the correlation between channels, while a low MI indicates more independent semantic information captured by different channels. Therefore, the experimental results demonstrate that improved performance is associated with decreased inter-channel correlation in the feature map.

2.3 Why Lower Inter-Channel Correlation is Better?

The reduced inter-channel correlation benefits our model in two aspects:

(1) **Cross-domain generalization**. Previous works [2, 51] indicate that a lower correlation between features implies reduced redundancy and enhanced generalizability. Intuitively, a lower inter-channel correlation demonstrates that channels capture patterns more independently, therefore each channel will capture patterns in the input image more uniformly, which means the mean magnitude of each channel across all images will be more uniform. Consistent with our intuition, [30] shows the channel bias problem affects the generalizability of few-shot methods, and it utilizes the Mean Magnitude of Channels (MMC) to visualize and measure the channel response in features, where effective few-shot methods might have a more uniform MMC curve in the testing set. Therefore, this means the reduced correlation also benefits our model by addressing the channel bias problem, as studied in [30].

Inspired by this, we visualize the MMC before and after applying the mask to filter frequency components on four target datasets. As shown in Figure 2, for FSS-1000, performance degrades after masking, with the curve steeper than the baseline. Conversely, for the other three target datasets, performance improves with the curve more uniform than the baseline after masking. This indicates the channel bias problem is also handled by frequency filtering, which benefits the model with more independent and diverse semantic patterns to represent target domains.

(2) **Exploring larger activated regions for segmentation**. To study why reducing inter-channel correlation through frequency filtering benefits the segmentation task, we visualize the heatmap of feature maps before and after filtering out specific frequency components. As shown in Figure 3(a), after filtering certain frequency components, the heatmap demonstrates expanded activation regions,

3

which results from more recognition patterns captured by independent channels. Since the segmentation task requires the model to pixel-wisely detect the whole semantic object, an expanded activation region means the model can better detect the entire object, instead of only focusing on the most discriminative parts.

2.4 Why Feature Disentanglement in the Frequency Domain?

In this subsection, we illustrate why the feature disentanglement is carried in the frequency domain.

Fourier Transform (FT)   FT transforms finite signals into complex-valued functions of frequency. For a single channel $f \in \mathbb{R}^{h \times w}$ of the feature map, the Fourier transform is formulated as:

$$F(u, v) = \frac{1}{wh} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} f(x, y) e^{-i2\pi\left(\frac{ux}{w} + \frac{vy}{h}\right)} \tag{1}$$

where i is imaginary unit, h and w are the height and the width of f. $f(x, y)$ is an element of f at spatial pixel (x, y), and $F(u, v)$ represents the Fourier coefficient at frequency component (u, v).

The process in which the spatial feature f is decomposed into the amplitude α and phase ρ is called spectral decomposition. The corresponding frequency feature F can be reassembled from amplitude α and phase ρ ($|F|$ is the modulus of F):

$$F = \alpha \cos(\rho) + i\alpha \sin(\rho) = \alpha \cdot e^{i\rho} \tag{2}$$

$$|F| = \sqrt{\alpha^2(\cos^2(\rho) + \sin^2(\rho))} = \sqrt{\alpha^2} = \alpha \tag{3}$$

Mathematical Derivation. To intuitively prove the correlation between phase differences and channel correlation within a feature map. For different channels of a feature map, representing distinct features, $F_1(m, n) = \alpha_1 e^{i\rho_1}$ and $F_2(m, n) = \alpha_2 e^{i\rho_2}$ are defined as the frequency domain representations of the same location in different channels. The correlation coefficient is calculated based on the following formula in the frequency domain:

$$r = \sum_{m=0}^{h-1} \sum_{n=0}^{w-1} \frac{F_1(m, n) F_2^*(m, n)}{\sqrt{|F_1(m, n)|^2 |F_2(m, n)|^2}} = \sum_{m=0}^{h-1} \sum_{n=0}^{w-1} r(m, n) \tag{4}$$

The $F_2^*(m, n)$ is the complex conjugate of $F_2(m, n)$, can be computed as:

$$F_2^*(m, n) = \alpha_2 \cos(\rho_2) - i\alpha_2 \sin(\rho_2) = \alpha_2 e^{-i\rho_2} \tag{5}$$

Substituting equations (2), (3), and (5) into equation (4) yields:

$$F_1(m, n) F_2^*(m, n) = \alpha_1 \alpha_2 e^{i(\rho_1 - \rho_2)} \tag{6}$$

$$|F_1(m, n)|^2 |F_2(m, n)|^2 = \alpha_1^2 \alpha_2^2 \tag{7}$$

from which we can further derive:

$$r(m, n) = \frac{\alpha_1 \alpha_2 e^{i(\rho_1 - \rho_2)}}{\sqrt{\alpha_1^2 \alpha_2^2}} = e^{i(\rho_1 - \rho_2)}, \quad \rho_1 - \rho_2 = \Delta\rho \in [0, \pi] \tag{8}$$

According to Euler's formula, we know that $e^{\pi i} = -1$ and $e^0 = 1$. From this derivation, we have proved that the correlation between features in the spatial domain can be translated into phase differences and amplitudes in the frequency domain. When the frequency components are identical, the following can be inferred: 1) $\Delta\rho = 0$, $r(m, n) = 1$ indicates perfect positive correlation; 2) $\Delta\rho = \pi$, $r(m, n) = -1$ indicates a perfect negative correlation. Therefore, when the phase differences of more corresponding points from different channels in the frequency domain aggregate around 0 or π, there is a higher correlation between the channels. For amplitude, when the phase differences are the same, the closer the amplitudes, the more similar the waveforms are, thus indicating a higher correlation between the channels.

4

(a) Feature Map Visualization   (b) Phase Difference Histograms
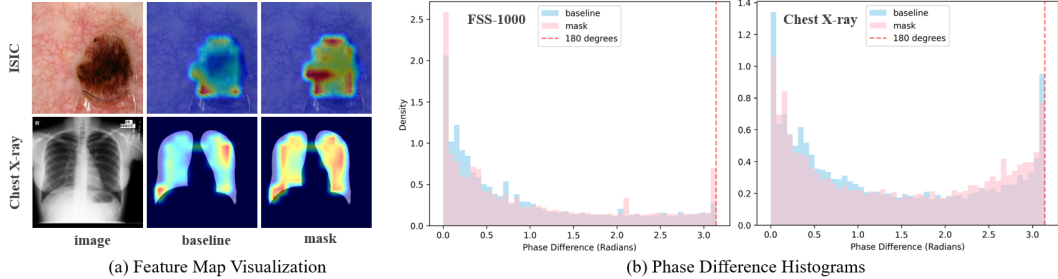
Figure 3: (a) After masking certain frequency components, the model's attention regions are enlarged with more patterns encompassed. (b) A higher concentration of phase differences at 0 and π indicates a higher correlation, so that on FSS-1000 the performance drops but on Chest X-ray it increases.

Experiments for derivation. To validate our derivations, we measured the phase difference histograms between channels, using amplitude as weights, to observe the relationship between phase differences among feature map channels. The phase difference and weight are as follows:

$$\Delta\rho = |\rho_1 - \rho_2|, \ \ weight = \frac{\alpha_1 \alpha_2}{|\alpha_1 - \alpha_2|} \tag{9}$$

The results shown in Figure 3(b), indicate that for FSS-1000, after masking certain frequency components, the inter-channel correlation increases. Correspondingly, phase differences between channels tend to cluster more around 0 and π. Conversely, for Chest X-ray, masking certain frequency components reduces inter-channel correlation, resulting in fewer phase differences clustering around 0 and π. The experimental results have validated the accuracy of our derived conclusions.

2.5 Conclusion and Discussion

Through mathematical derivation and experiments, we demonstrate that manipulating the frequency can reduce inter-channel correlation in feature maps. Each channel of the feature map represents a distinct pattern, and a lower inter-channel correlation implies a higher degree of channel disentanglement, leading to more independent and diverse semantic patterns for each feature. This benefits the model with 1) alleviation of channel bias, boosting model robustness on target domains; and 2) larger activated regions for segmentation, therefore a simple frequency filtering operation can significantly improve performance for the CD-FSS task.

Based on the above analysis, we can draw the following insights: 1) The aforementioned mask operates at the input level, but fundamentally affects the feature map's channel correlation. Therefore, we can directly apply mask operations to the frequency domain of each channel in the feature map; 2) Different domains require filtering different components. The aforementioned mask manually filters different frequencies based on the target domain, but the mask can be made adaptive; 3) The aforementioned mask does not perform well on FSS-1000. We believe this may be due to the overly coarse high-low frequency division. A finer frequency division can be designed, dividing the frequency into $h \times w$ parts (where h and w are the spatial dimensions of the feature map).

3 Method

Our method consists of two major steps, i.e., 1) amplitude-phase masker is proposed to reduce feature correlation, and obtain more accurate and generalized feature maps; 2) adaptive channel phase attention is proposed to select features that benefit the current instance and align the feature spaces of support and query. Our modules do not require source-domain training and can be directly integrated into during target-domain fine-tuning. The overall framework of our approach is shown in Figure 4.

3.1 Amplitude-Phase Masker

Amplitude-Phase Masker(APM) is a model-agnostic module that filters out negative frequency components at the feature level within feature maps. Through mathematical derivation, it is shown that APM accomplishes feature disentanglement. Consequently, this leads to a feature map that is more robust, generalizable, and provides broader and more accurate representations.

In our work, we utilize a fixed encoder, trained on the source domain, to extract feature maps $\mathcal{F} \in \mathbb{R}^{c \times h \times w}$, where c, h and w represent channels, height, and width. We then apply the Fast

5
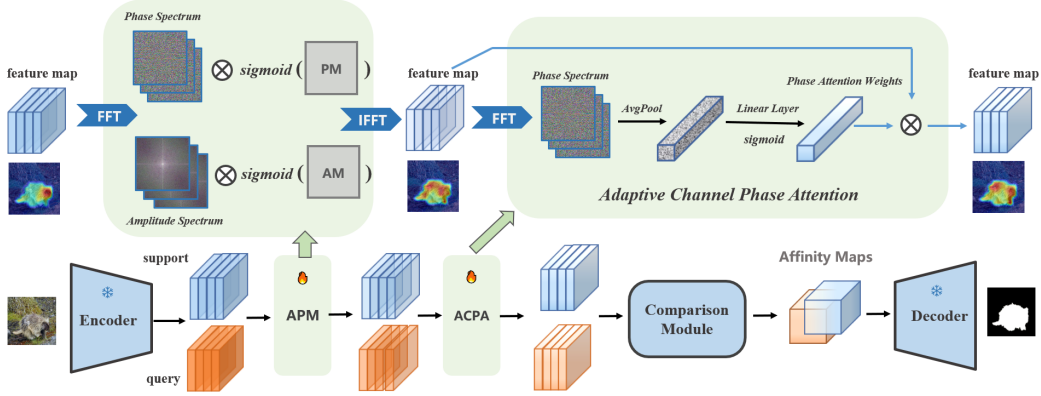
Figure 4: Overview of our method in a 1-shot example. After obtaining the feature map, APM is introduced to adaptively filter certain frequency components based on different domains, facilitating feature disentanglement to achieve more generalizable representations. Additionally, we propose ACPA to encourage the model to focus on more effective features while aligning the feature space of the support and query images. The internal structure of APM and ACPA is highlighted in green.

Fourier Transformation (FFT) to convert these feature maps from the spatial domain into the frequency domain, further decomposing them into phase spectrum $\mathcal{P}$ and amplitude spectrum $\mathcal{A}$:

$$\mathcal{A}e^{i\mathcal{P}} = FFT(\mathcal{F}) \tag{10}$$

Then, the phase and amplitude obtained from the FFT are each subjected to a Hadamard product with their respective sigmoid-activated phase mask (PM) $\mathcal{M}_p$ and amplitude mask (AM) $\mathcal{M}_a$. This operation effectively filters out negative components from both the phase and amplitude:

$$\mathcal{P}_{enh} = Sigmoid(\mathcal{M}_p) \otimes \mathcal{P}$$
$$\mathcal{A}_{enh} = Sigmoid(\mathcal{M}_a) \otimes \mathcal{A} \tag{11}$$

where $\otimes$ indicates the element-wise multiplication, $\mathcal{P}_{enh}$ and $\mathcal{A}_{enh}$ denotes the enhanced phase and amplitude, respectively. For each task, the APM is initialized with all ones, where $Sigmoid(\mathcal{M}_*) \in [0, 1]$. Here, 1 allows complete passage of frequency components, while 0 results in their total filtration. The original APM (APM-S) is a lightweight module, configured as an $h \times w$ matrix that matches the height and width of the feature map. We also offer a variant APM (APM-M) that expands the dimensions to $c \times h \times w$, in alignment with the feature map's dimensions.

The filtered phase and amplitude components are recombined and transformed back into the spatial domain using the Inverse Fast Fourier Transform (IFFT) to produce the enhanced feature map:

$$\mathcal{F}_{enh} = IFFT(\mathcal{A}_{enh}e^{i\mathcal{P}_{enh}}) \tag{12}$$

We iterate and optimize the APM using the support and its corresponding labels. After the APM process, the model generates a feature map that is more accurate and generalizable. This feature map is then fed into the subsequent Adaptive Channel Phase Attention module for further optimization.

3.2 Adaptive Channel Phase Attention

Adaptive Channel Phase Attention(ACPA) can be seen as a process of feature selection. Building on the APM-optimized feature map, ACPA encourages the model to focus on more effective channels (features) while aligning the feature spaces of the support and query. Its underlying insight is that amplitude and phase are considered as style and content, respectively. Therefore, the phase can be seen as an invariant representation, with consistent phase elements across both support and query.

For the enhanced support feature maps $\mathcal{F}_{enh}^{sup}$, a support mask $M^s \in \{0,1\}^{H \times W}$ is applied to discard irrelevant activations:

$$\mathcal{F}_{mask}^{sup} = \mathcal{F}_{enh}^{sup} \otimes M^s \tag{13}$$

Subsequent operations input to ACPA are consistent with those applied to the query feature maps. Therefore, we uniformly refer to the feature map fed into ACPA as $\mathcal{F}_{enh}$. We adopt a SE block following SENet [19] as our channel attention module, denoted as SE:

$$\mathcal{W}_{phase} = SE(\mathcal{P}_{enh}) \tag{14}$$

6

Table 2: Mean-IoU of 1-shot and 5-shot results on the CD-FSS benchmark. The best and second-best results are in bold and underlined, respectively. * denotes the model implemented by ourselves. APM-S is an $1 \times h \times w$ matrix, while APM-M (more parameters) expands to $c \times h \times w$.

Method	FSS-1000		Deepglobe		ISIC		Chest X-ray		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PGNet [45]	62.42	62.74	10.73	12.36	21.86	21.25	33.95	27.96	32.24	31.08
PANet [41]	69.15	71.68	36.55	**45.43**	25.29	33.99	57.75	69.31	47.19	55.10
CaNet [46]	70.67	72.03	22.32	23.07	25.16	28.22	28.35	28.62	36.63	37.99
RPMMs [43]	65.12	67.06	12.99	13.47	18.02	20.04	30.11	30.82	31.56	32.85
PFENet [37]	70.87	70.52	16.88	18.01	23.50	23.83	27.22	27.57	34.62	34.98
RePRI [5]	70.96	74.23	25.03	27.41	23.27	26.23	65.08	65.48	46.09	48.34
HSNet [31]	77.53	80.99	29.65	35.08	31.20	35.10	51.88	54.36	47.57	51.38
HSNet* [31]	77.54	80.21	33.19	36.46	32.65	35.09	47.34	48.63	47.68	50.10
PATNet [24]	<u>78.59</u>	<u>81.23</u>	37.89	42.97	41.16	**53.58**	66.61	70.20	56.06	61.99
Ours (APM-S)	78.25	80.29	<u>40.77</u>	44.85	<u>41.48</u>	49.39	<u>75.22</u>	<u>76.89</u>	<u>58.93</u>	<u>62.86</u>
Ours (APM-M)	**79.29**	**81.83**	**40.86**	<u>44.92</u>	**41.71**	<u>51.16</u>	**78.25**	**82.81**	**60.03**	**65.18**

where $\mathcal{W}_{phase} \in \mathbb{R}^{c \times 1 \times 1}$ is phase attention weight, $\mathcal{P}_{enh}$ is the phase of $\mathcal{F}_{enh}$. Then we apply the phase attention weights to the feature map $\mathcal{F}_{enh}$ to obtain the final feature map $\mathcal{F}_{final}$:

$$\mathcal{F}_{final} = \zeta_l(\mathcal{W}_{phase}) \otimes \mathcal{F}_{enh} \tag{15}$$

where $\otimes$ is the Hardmard product and $\zeta_l(*)$ extends the weight to match the dimension of the feature map by expanding along the spatial dimension, i.e., $\zeta_l : \mathbb{R}^{c \times 1 \times 1} \to \mathbb{R}^{c \times h \times w}$.

Finally, a pair of query feature maps $\mathcal{F}_{final}^{qry}$ and support feature maps $\mathcal{F}_{final}^{sup}$ are fed into comparison module forms affinity maps $C \in \mathbb{R}^{h \times w \times h \times w}$ using cosine similarity:

$$C(m, n) = ReLU\left(\frac{\mathcal{F}_{final}^{qry}(m) \cdot \mathcal{F}_{final}^{sup}(n)}{\|\mathcal{F}_{final}^{qry}(m)\|\|\mathcal{F}_{final}^{sup}(n)\|}\right) \tag{16}$$

where m, n denote 2D spatial positions of feature maps $\mathcal{F}_{final}^{qry}$ and $\mathcal{F}_{final}^{sup}$ respectively. Then, the $C(m, n)$ is fed into the decoder to obtain segmentation results, as shown in Figure 4.

4 Experiments

4.1 Datasets

We utilize the benchmark established by PATNet [24] and adopt the same data preprocessing methods. For training, our source domain is the PASCAL-5^i dataset [35], an extended version of PASCAL VOC 2012 [13] enhanced with additional annotations from the SDS dataset. For evaluation, our target domains include FSS-1000 [26], Deepglobe [10], ISIC2018 [9, 38], and the Chest X-ray datasets [6, 20]. See Appendix A.1 for more details about datasets.

4.2 Implementation Details

We employ ResNet-50 [18] as our encoder, initialized with weights pre-trained on ImageNet [34]. The training manner is consistent with our baseline model HSNet [31]. To optimize memory usage and speed up training, the spatial sizes of both support and query images are set to 400×400. The model is trained using the Adam [21] optimizer with a learning rate of 1e-3.

During the adaptation stage, the model initially predicts the support mask and then uses the corresponding label to optimize the APM and ACPA through CE loss. The adaption stage of the APM and the ACPA leverage feature maps from conv5_x whose channel dimensions are 2048 and spatial size is 13×13, is performed using the Adam optimizer, with learning rates set at 0.1 for Chest X-ray, 0.01 for FSS-1000 and ISIC, and 1e-5 for Deepglobe. Each task undergoes a total of 60 iterations.

4.3 Comparison with State-of-the-Art Works

In Table 2, we compare our method with several state-of-the-art few-shot semantic segmentation approaches on the benchmark introduced by PATNet [24]. Our results show a significant improvement in cross-domain semantic segmentation for both 1-shot and 5-shot tasks. Specifically, APM-S exceeds the performance of the state-of-the-art PATNet, based on ResNet-50, by 2.87% and 0.87% in average

7
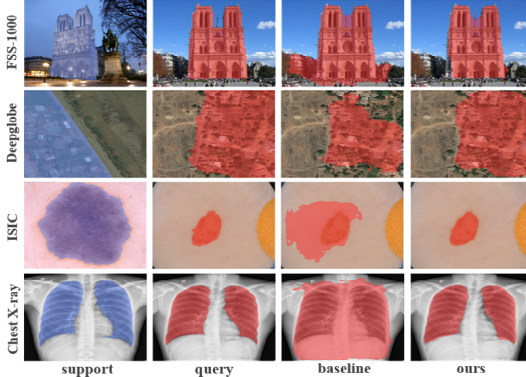
Figure 5: Qualitative results of our model.

Table 3: Ablation study on various designs

APM-S	APM-M	ACPA	1-shot	5-shot
			47.68	50.10
✓			57.77	61.39
✓		✓	**58.93**	**62.86**
	✓		59.13	63.53
	✓	✓	**60.03**	**65.18**

Table 4: APM-S implemented in the transformer architecture.

	FSS	Deepglobe	ISIC	Chest	Average
perSAM [48]	79.65	33.39	21.27	31.12	41.35
FPTrans [47]	78.90	38.29	47.60	78.92	60.93
FPTrans+ours	**79.84**	**39.78**	**50.37**	**79.29**	**62.32**

mIoU for the 1-shot and 5-shot settings, respectively. APM-M outperforms the state-of-the-art by 3.97% and 3.19%. Additionally, Figure 5 presents qualitative results of our method in 1-way 1-shot segmentation, highlighting a substantial enhancement in generalization across large domain gaps while maintaining comparable accuracy with similar domain shifts.

4.4 Ablation Study

Effectiveness of each module. We evaluated each proposed module in both 1-shot and 5-shot settings to assess their effectiveness. As detailed in Table 3, introducing APM-S increased the average mIoU by 10.09% for 1-shot and 11.29% for 5-shot. Adding ACPA further enhanced the mIoU by 1.16% and 1.47%, respectively. Additionally, APM-M, a variant of APM with more parameters, when combined with ACPA, increased the average mIoU by 12.35% for 1-shot and 15.08% for 5-shot.

Model-agnostic method. We implemented our method for FPTrans [47], which employs the Vision Transformer (ViT) [12] as its encoder. As shown in Table 4, our approach can also effectively enhance the performance of models based on the transformer architecture. Moreover, the new large-scale SAM [22] model has significantly advanced image segmentation, showcasing impressive zero-shot capabilities. However, SAM is not suited for cross-domain few-shot segmentation. Thus, we evaluate PerSAM [48] to compare our method with the SAM-based approach. The result shows that our method performs much better than PerSAM in cross-domain few-shot segmentation.

Comparison with other method. To demonstrate the effectiveness of our method, we compared it with full parameter fine-tuning and feature disentangling method, as shown in Table 5. Compared to full parameter fine-tuning, our method uses fewer parameters and achieves better performance. For spatial domain feature disentangling, we added a mutual information loss to the baseline model during training to encourage each channel of the feature map to learn independent representations. Our method significantly outperforms this approach. Intuitively, feature disentanglement in the frequency domain offers finer granularity and global representation, which is more beneficial for segmentation tasks compared to the local representation in the spatial domain.

4.5 APM: Feature Disentanglement via Frequency Operations

Reduce inter-channel correlation. As shown in Table 6, we validated APM's ability to reduce inter-feature correlation and improve generalization performance by calculating the mutual information between feature map channels. Compared to Table 1, this result demonstrates that the adaptive feature-level approach is more effective than the input-level masking, further reducing inter-feature correlation. Furthermore, we plotted the cumulative distribution function (CDF) of inter-channel correlations in the feature maps, as shown in Figure 6. It can be observed that with the inclusion of APM, the CDF curve shifts to the left, indicating a decrease in inter-channel correlations.

Table 5: Compare our method with fine-tuning and spatial domain feature disentangle method.

	FSS	Deepglobe	ISIC	Chest	Average
baseline	77.54	33.19	32.65	47.34	47.68
baseline + ft	78.06	33.54	33.28	74.16	54.76
baseline+MI Loss	78.90	32.28	32.53	52.12	48.76
ours(APM-M)	**79.29**	**40.86**	**41.71**	**78.25**	**60.03**

Table 6: Verify the effectiveness of the APM by the mean mutual information (MI). (w/o ACPA)

APM-S	FSS	Deepglobe	ISIC	Chest
1-shot mIoU	77.98↑	40.74↑	38.79↑	73.55↑
support MI	1.3501↓	1.2761↓	1.3139↓	1.2610↓
query MI	1.3509↓	1.2794↓	1.3121↓	1.2620↓

8

Figure 6: Cumulative distribution function (CDF) of inter-channel correlations. After passing through APM, the CDF curve shifts to the left, indicating a decrease in inter-channel correlations.

More independent semantic representations. As shown in Table 7 (Left), we visualized the heatmaps of feature maps processed by APM and those not processed by APM. It is evident that after applying APM, the model focuses on more different features. For example, in the first column, the baseline only highlights the swan's head, whereas APM makes each feature representation more independent, allowing the model to attend to various features of the swan.

4.6 ACPA: Aligning Task-Relevant Features and Feature Spaces

ACPA can be seen as feature selection, enabling the model to focus on features that are more effective for the current task while aligning the feature spaces of the support and query feature maps. As shown in Table 7 (Left), after APM disentangles the features and produces a more broadly represented feature map, ACPA selects features that are more effective and discriminative for the current task. For example, in the first row, ACPA selects the swan's wings and head. In the second column, it selects the bird's head, tail, and feet. Furthermore, we measured the CKA (Centered Kernel Alignment) to calculate the distance between the support feature map and the query feature map, validating that ACPA aligns the support and query feature spaces. CKA is proposed to measure both intra-domain and inter-domain distances [52]; the smaller the CKA value, the closer the feature spaces. As shown in Table 7 (Right), after applying APM, the distance between support and query is reduced, and with the addition of ACPA, the support and query feature spaces are further aligned.

APM-S	APM-M	ACPA	FSS	Deepglobe	ISIC	Chest
			0.3591	0.2691	0.2494	0.5848
✓			0.3481	0.2678	0.2433	0.5025
✓		✓	0.2907	0.2676	0.2032	0.3628
	✓		0.3293	0.2675	0.2310	0.4635
	✓	✓	**0.2883**	**0.2674**	**0.1811**	**0.3986**

Table 7: (Left) Feature map visualizations show: 1) APM achieves feature disentanglement. 2) ACPA encourages the model to focus on more effective features. (Right) Verify the effectiveness of the ACPA in aligning the support and query feature spaces by CKA measure.

4.7 Comparison with Domain Transfer Methods

We compare our method against traditional frequency-based and correlation-based approaches to validate our method's effectiveness. For a fair comparison, all methods are implemented on the baseline HSNet [31] and then evaluated under the 1-shot setting on the CD-FSS benchmark.

Frequency-based method DFF [27] preserves frequency information beneficial for generalization. GFNet [32] replaces self-attention layer with global frequency filter layer. ARP [7] introduces Amplitude-Phase Recombination via amplitude transformation. DAC [23] proposes a normalization method that removes style (amplitude) while preserving content (phase) through spectral decomposition. Although these methods improve generalization, they fall short in addressing large domain gaps. Our method requires no source domain training. It adaptively masks harmful components for the target domain at the feature level. By treating amplitude and phase separately, we exploit phase invariance to design a channel attention module that handles intra-class variations. As shown in Table 8 our method outperforms existing frequency-based approaches on the CD-FSS task.

9

Table 8: Compare our method to previous frequency-based methods under 1-shot setting.

	FSS	Deepglobe	ISIC	Chest	Average
baseline [31]	77.54	33.19	32.65	47.34	47.68
DFF [27]	78.18	32.16	35.71	60.29	51.59
GFNet [32]	76.86	32.23	33.95	53.12	49.04
ARP [7]	78.83	35.06	35.61	59.83	52.33
DAC [23]	78.83	35.98	36.02	57.66	51.98
ours(APM-M)	**79.29**	**40.86**	**41.71**	**78.25**	**60.03**

Table 9: Compare our method to other reducing correlation approaches under 1-shot setting.

	FSS	Deepglobe	ISIC	Chest	Average
baseline [31]	77.54	33.19	32.65	47.34	47.68
MMC(Simple) [30]	77.48	34.70	34.32	48.74	48.81
MMC(Oracle) [30]	77.45	35.12	34.59	50.27	49.36
SRIP [1]	78.13	34.61	34.05	50.58	49.34
baseline+whitening	77.92	33.22	32.98	50.89	48.75
ours(APM-M)	**79.29**	**40.86**	**41.71**	**78.25**	**60.03**

Correlation-based method   For methods that directly constrain the model (orthogonality, whitening): the few-shot setting means limited sample size, and existing models have a large number of parameters. Directly adjusting the model with constraints using such small datasets is not effective and even can lead to negative optimization. As seen in Table 9, the performance of orthogonality constraints(SRIP [1]) and whitening is not satisfactory. For feature transformation/augmentation methods like MMC [30]: the stability is not guaranteed because they use specific feature transformation functions. Due to the domain gap, a transformation method effective for one domain may not be effective for others. In contrast, our method has the advantages of being 1) lightweight (allowing for quick adaptation in the few-shot setting) and 2) stable and robust (with adaptive adjustments for different target domains). These benefits are well reflected in the performance results.

5   Related Work

Few-shot learning   Few-shot learning aims to build robust representations for new concepts with limited annotated examples. Existing approaches typically fall into three categories: metric learning [36, 39], optimization-based methods [14, 33] and graph-based methods [15, 28]. Recently, cross-domain few-shot learning has gained attention due to disparities in both data distribution and label space between meta-testing and meta-training stages. BSCD-FSL [16] introduces a challenging benchmark for cross-domain few-shot learning, featuring a substantial domain gap between the source and target domains. It covers several target domains with varying similarities to natural images.

Few-shot semantic segmentation   Few-shot semantic segmentation aims to segment unseen classes in query images with only a few annotated samples. OSLSM [35] is the first two-branch FSS model. Following this, PL [11] introduces a prototype learning paradigm utilizing cosine similarity between pixels and prototypes. SG-One [49] adopts masked average pooling (MAP) to optimize the extraction of support features. Recently, many FSS methods have emerged in the research community, such as RPMMs [43], PFENet [37], ASGNet [25], and HSNet [31]. HSNet employs efficient 4D convolutions on multi-level feature correlations, serving as the baseline for our work. However, these methods primarily address segmenting novel classes within the same domain and struggle with generalization across disparate domains due to significant feature distribution disparities. Bridging this substantial domain gap, particularly with limited labeled data, remains a formidable challenge.

6   Conclusion

In this paper, we delve into the phenomenon that filtering specific frequency components based on different domains significantly improves performance, providing an interpretation through experiment and mathematical derivation. Building on our interpretation, we propose the APM, a feature-level frequency component mask designed to enhance the generalization of feature map representations. Further, we introduced ACPA. Based on the APM-optimized feature map, the ACPA encourages the model to focus on more effective features while aligning the feature spaces of the support and query. Experimental results demonstrate the approach's effectiveness in reducing domain gaps.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62206102, 62436003, 62376103 and 62302184; the Science and Technology Support Program of Hubei Province under grant 2022BAA046; Hubei Science and Technology Talent Service Project under grant 2024DJC078; and Ant Group through CCF-Ant Research Fund.

10

References

[1] N Bansal, X Chen, and Z Wang. Can we gain more from orthogonality regularizations in training deep cnns? arxiv 2018. *arXiv preprint arXiv:1810.09102*.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[5] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13979–13988, 2021.

[6] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

[7] 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*, pages 458–467, 2021.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[9] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[10] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.

[11] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[15] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

11

[16] Yunhui Guo, Noel CF Codella, Leonid Karlinsky, John R Smith, Tajana Rosing, and Rogerio Feris. A new benchmark for evaluation of cross-domain few-shot learning. *arXiv preprint arXiv:1912.07200*, 2019.

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

[18] 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*, pages 770–778, 2016.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[20] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[23] Sangrok Lee, Jongseong Bae, and Ha Young Kim. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11776–11785, 2023.

[24] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 73–90. Springer, 2022.

[25] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8334–8343, 2021.

[26] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020.

[27] Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11797–11807, 2023.

[28] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[30] Xu Luo, Jing Xu, and Zenglin Xu. Channel importance matters in few-shot image classification. In *International conference on machine learning*, pages 14542–14559. PMLR, 2022.

[31] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021.

12

[32] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

[33] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

[34] 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*, 115:211–252, 2015.

[35] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[37] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020.

[38] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[40] 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*, pages 8684–8694, 2020.

[41] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.

[42] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

[43] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 763–778. Springer, 2020.

[44] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

[45] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019.

[46] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5217–5226, 2019.

[47] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in neural information processing systems*, 35:6575–6588, 2022.

[48] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

13

[49] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE transactions on cybernetics*, 50(9):3855–3865, 2020.

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[51] Haichen Zhou, Yixiong Zou, Ruixuan Li, Yuhua Li, and Kui Xiao. Delve into base-novel confusion: Redundancy exploration for few-shot class-incremental learning. *arXiv preprint arXiv:2405.04918*, 2024.

[52] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Advances in neural information processing systems*, 35:27267–27279, 2022.

14

A   Appendix / supplemental material

A.1   More Dataset Details
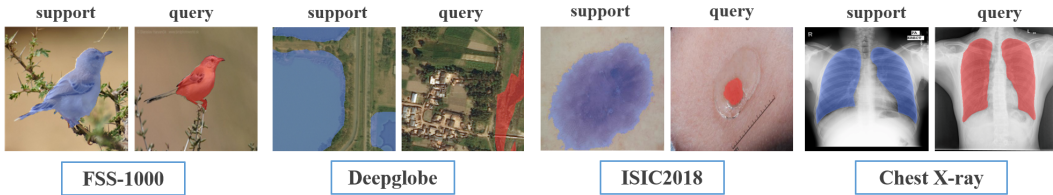
Figure 7: Examples of segmentation for four target datasets.

Our experimental setup is grounded in the benchmark established by PATNet [24]. Fig. 7 presents an example of segmentation for four target datasets. Further details follow:

PASCAL-5^i [35] extends the PASCAL VOC 2012 [13] by integrating additional annotations from the SDS dataset [17]. Utilizing PASCAL-5^i as our source domain for training, we then evaluate the models' performance across four target datasets.

FSS-1000 [26] is a few-shot segmentation dataset comprising 1000 natural image categories, with each category containing 10 samples. In our experiment, we adhere to the official split for semantic segmentation and report results on the designated testing set, which encompasses 240 classes and 2,400 images. We consider FSS-1000 as our designated target domain for testing.

Deepglobe [10] consists of satellite images annotated densely at the pixel level across 7 categories: urban, agriculture, rangeland, forest, water, barren, and unknown. Since ground-truth labels are available only in the training set, we utilize the official training dataset comprising 803 images to showcase our results. We designate it as our testing target domain and follow the same processing approach as PATNet.

ISIC2018 [9, 38] is designed specifically for skin cancer screening and comprises images of lesions, with each image depicting exactly one primary lesion. Adhering to the guidelines established by PATNet, we process and utilize the dataset, considering ISIC2018 as our target domain for testing.

Chest X-ray [6, 20] is tailored for Tuberculosis diagnosis, comprising 566 images with a resolution of 4020 × 4892 pixels. These images depict cases from 58 Tuberculosis patients and 80 individuals with normal conditions. A common approach to handling large image sizes involves resizing them to 1024 × 1024 pixels.

A.2   APM Reduces Inter-Channel Correlation by Frequency Domain Mask
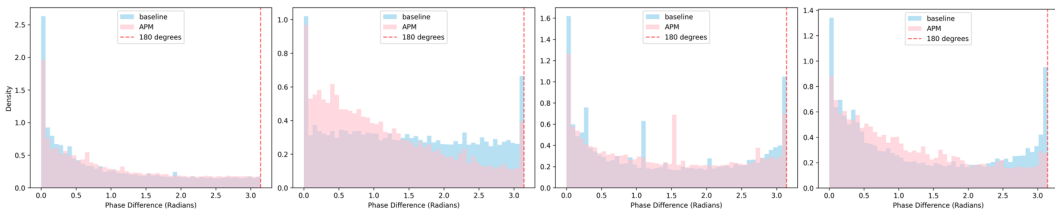
Figure 8: Histogram of phase differences (weighted by amplitude) between channels in the feature maps before and after APM.

As shown in Figure 8, we present histograms of the phase differences between channels in the feature maps (weighted by amplitude) before and after adding the APM module. After applying APM, the phase differences between channels concentrated around 0 and π are reduced, which aligns with our mathematical derivation in the main text. The APM decreases the correlation between feature map channels by altering phase and amplitude, thereby enhancing the independence of their semantic representations. Additionally, due to the use of finer-grained frequency domain partitioning, APM performs well on FSS-1000 compared to the simple high-low frequency partitioning at the input level.

15
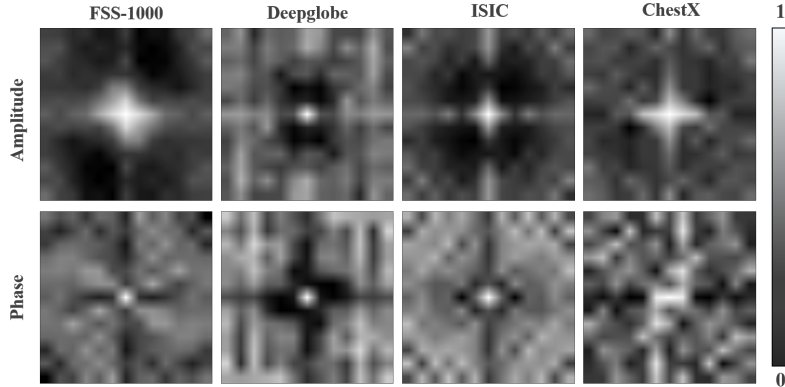
Figure 9: The visualization of the frequency components filtered by the masker.

A.3 Analyze the Frequency Components Filtered by the APM

The visualization results in Figure 9 show the average frequency components filtered by the amplitude masker and phase masker across different domains. The center represents low frequencies, while the periphery represents high frequencies. White indicates a value of 1, meaning the frequency component passes through, and black indicates a value of 0, meaning the frequency component is filtered out. For FSS, the amplitude masker (AM) primarily filters out mid-to-high-frequency components, while the phase masker (PM) mainly retains mid-frequency components. For DeepGlobe, both AM and PM retain more mid-to-high frequencies. For ISIC, AM filters out more mid-to-high frequencies, retaining low frequencies, whereas PM retains relatively more mid frequencies. For ChestX, AM mainly retains low-frequency components, while PM filters out frequencies across the spectrum, retaining relatively more low-to-mid frequencies. These results align well with the patterns observed in Figure 1 of our main text. It is evident that for different targets, AM and PM dynamically and adaptively filter different frequency components, selecting those more beneficial for the current domain. Additionally, the advantageous amplitude and phase frequency components vary across different target domains, underscoring the necessity of considering amplitude and phase separately.

A.4 Detailed Ablation Study Results

In the main text, we demonstrate the effectiveness of our various designs by presenting the average mIoU. Here, as shown in Table 10, we provide detailed results on each target dataset.

Table 10: Detailed ablation study results of various designs (Backbone: ResNet-50).

APM-S	APM-M	ACPA	FSS-1000		Deepglobe		ISIC		Chest X-ray		Average	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
			77.54	80.21	33.19	36.46	32.65	35.09	47.34	48.63	47.68	50.10
✓			77.98	79.85	40.74	44.80	38.79	44.16	73.55	76.73	57.77	61.39
✓		✓	78.25	80.29	40.77	44.85	41.48	49.39	75.22	76.89	58.93	62.86
	✓		78.98	81.21	40.81	44.82	38.99	45.49	77.73	82.60	59.13	63.53
	✓	✓	79.29	81.83	40.86	44.92	41.71	51.16	78.25	82.81	60.03	65.18

A.5 Broader Impact

Our research delves into the phenomenon that filtering specific frequency components based on different domains significantly improves performance, providing an interpretation. Furthermore, we demonstrated the relationship between frequency components and inter-channel correlation through mathematical derivation. Building on our interpretation and derivation, we propose the APM, a feature-level frequency component mask designed to enhance the generalization of feature map representations. Further, we introduced Adaptive Channel Phase Attention (ACPA). Based on the APM-optimized feature map, the ACPA encourages the model to focus on more effective features while aligning the feature spaces of the support and query. Experimental results demonstrate the effectiveness of our approach significantly enhances the model's cross-domain transferability. This work is applicable not only to CDFSS but also to other areas like domain generalization and domain adaptation. Future research will aim to broaden our evaluations to encompass a wider range of target domains, enhancing our understanding of their performance in various real-world scenarios.

16

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

17

Justification: We have provided the full set of assumptions and a complete (and correct) proof for each theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have reported the detailed settings and hyper-parameters. We will release our codes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

18

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have reported the detailed settings and hyper-parameters. We will release our codes.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have reported the detailed settings and hyper-parameters. We will release our codes.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Experiments are conducted based on the code released by HSNet (ICCV'21) and benchmark proposed by PATNet (ECCV'22), which specified the random seed for re-implementation and comparison. Therefore, we follow the same random seed to conduct experiments and compare with other methods without the error bar.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

19

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

    Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

    Answer: [Yes]

    Justification: We have reported the computer resources in the implementation details.

    Guidelines:

    - The answer NA means that the paper does not include experiments.
    - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
    - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
    - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

    Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

    Answer: [Yes]

    Justification: We conform, in every respect, with the NeurIPS Code of Ethics.

    Guidelines:

    - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
    - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
    - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the broader impact in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

20

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not have the risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have respected the license of assets used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

21

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not have new assets in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not include human subjects information.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not include human subjects information.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

22