

Supplementary Materials: The Name of the Title is Hope

Anonymous Authors

A EXPERIMENTAL SETTINGS

A.1 Datasets

Benchmark datasets. To demonstrate the comprehensive adaptability of our framework to visual recognition domains in a black-box setting, we conducted experiments on 14 benchmark datasets following the protocol of [13, 19, 20]. Benchmark datasets covered a spectrum of tasks, including generic objects recognition on ImageNet [5] and Caltech101 [6], fine-grained categories classification on OxfordPets [14], StanfordCars [10], Flowers102 [12], and Food101[2], scene recognition on SUN397 [18], satellite imagery recognition on EuroSAT [7] and RESISC45 [3], object counting on CLEVR [9], and specific classification tasks such as texture, digit, and action using DTD [4], SVHN [11], and UCF101[16].

Synthetic datasets. To evaluate the robustness of our framework against distribution shift and adversarial noise, we measured the performance on Biased MNIST and Loc-MNIST, which are also introduced in Oh et al. [13]. Biased MNIST, derived from handwritten digit images in MNIST by altering background and font colors, is designed to assess model stability under distribution shift. The authors changed the background color of digit images, aiming for a high correlation ρ between label and background in the training dataset, while deliberately reducing the correlation $(1 - \rho)$ in the validation dataset.

Loc-MNIST is a dataset comprising images where two digits are drawn on the edge and center of a completely black background for the task of classifying the edge digit while ignoring the center digit. Testing on this dataset is associated with the capability to withstand adversarial noise and ability to recognize digits regardless of location.

A.2 Baseline Methods

Zero-shot classification (ZS). CLIP [15] is the most renowned pre-trained vision recognition model, demonstrating remarkable performance across a range of tasks, from image classification to visual reasoning, through its versatile design. By unifying image and text feature space, CLIP breaks the limitations of closed-set problem definition in conventional supervised learning, enabling classification for undefined visual concept, known as zero-shot classification (ZS).

Black-box adversarial reprogramming (BAR). BAR [17] introduces adversarial program to repurpose publicly undisclosed pre-trained models for various vision tasks, including medical imaging. Assessed across three clinical imaging tasks, BAR demonstrated performance comparable to the white-box adversarial programming, surpassing both the state-of-the-art methods of each dataset and the commonly employed fine-tuning approach.

Visual prompting (VP). VP [1] serves as a transfer learning approach by adding a learnable visual prompt in the form of padding

to images, introducing the concept of text prompts into vision tasks to enhance adaptation performance. Through diverse experiments on pre-trained models and datasets, the authors demonstrated that visual prompts can yield results competitive with linear probes.

BlackVIP. BlackVIP [13] proposes a novel input-dependent visual prompt design within black-box settings. Specifically, they leverage a frozen image encoder to extract image features, then employ a decoder to generate a visual prompt matching the image size; later, this is added to images at the pixel level. In addition, SPSA-GC is also proposed as an enhanced zeroth-order optimization method to mitigate unstable optimization.

A.3 Implementation Details

Decoder architecture. As illustrated in Fig. 1(a) in the main paper, our approach generates visual prompts without the inclusion of an additional pre-trained proficient encoder. The decoder comprises a total of 5 convolutional blocks, first taking the learnable trigger vector ϕ^{v_1} as input, then taking another learnable trigger vector ϕ^{v_2} as additional input in the middle of the decoder operation, and finally outputs both the spatial-domain VP V_s and the frequency-domain VP V_f . In detail, spatial V_s is generated as ϕ^{v_1} passes through four convolutional blocks within the decoder, while V_f is generated by concatenating the features from the third convolutional block with ϕ^{v_2} and passing through one remaining convolutional block. V_s has the same size as the resized image and V_f has the same size as the original image. In Algorithm 1, we describe the pseudo-code detailing the application process of the spatial-domain VP V_s and frequency-domain VP V_f .

Algorithm 1 PyTorch-style Pseudo Code for Prompting Image

```
# image: an input image with shape of 224 x 224 x 3
# Decoder: spatial-frequency hybrid prompter
# t_1, t_2: trigger vectors to generate spatial and frequency VP
# dct, idct: DCT and inverse-DCT operation respectively
# scale_factor: learnable scale parameter in frequency domain
# ZeroPad: function that appends zero padding with a specific margin
# Mask: function that applies a center-oriented square mask with a specific length

# Generate VPs with decoder then apply additional operations
# to prepare to combine with the input image.
spatial_vp, frequency_vp = Decoder(t_1, t_2)
frequency_vp = ZeroPad(dct(frequency_vp), margin=(0, p, 0, p))
spatial_vp = Mask(spatial_vp, length=h)

# Resize and apply DCT to image
freq_domain_image = dct(resize(image))

# Add frequency VP to trasformed image
freq_combined = freq_domain_image + sigmoid(scale_factor) * frequency_vp
spatial_domain_image = idct(freq_combined)

# Combine the IDCT image with spatial visual prompt
final_prompted_image = spatial_domain_image + spatial_vp
```

Intra-class Relation Loss. Previous work [8] propose the intra-class relation loss in context of knowledge distillation with aim of preserving the relational dynamics within each class. To this end, the authors formulate the loss function to maximize the column-wise correlation between a source prediction matrix \mathbb{Y}^s and a target

one as following:

$$\mathcal{L}_{\text{intra}}^{KD} = \frac{1}{C} \sum_{j=1}^C d_p(Y_{:,j}^{(s)}, Y_{:,j}^{(t)}), \quad (1)$$

where C and d_p is a number of class and Pearson's distance respectively. In our scenario, we adapt this formulation to enhance the alignment between original prediction probabilities and refined prediction probabilities matrices: $T^{(s)} = \{P_\theta(V_\phi(\mathbf{X}_i))\}_{i=1}^B$, $T^{(r)} = \{P_a(V_\phi(\mathbf{X}_i))\}_{i=1}^B \in \mathbb{R}^{B \times K}$ in batch size B . This can be written as:

$$\mathcal{L}_{\text{intra}} = \frac{1}{K} \sum_{j=1}^K d_p(T_{:,j}^{(s)}, T_{:,j}^{(r)}). \quad (2)$$

The relation loss contributes to maintain the relative order or preferences among instances within the same class, thereby enhancing performance.

A.4 Additional Experiments

Ablation study for pre-trained model backbones. To explore the adaptability of our approach, we adjusted the underlying architecture of the pre-trained target model. Unlike the BAR and VP methods, which do not enhance the zero-shot visual recognition capabilities of CNN-based backbones such as ResNet-50 (RN50) and ResNet-101 (RN101), our technique consistently achieves significant performance improvements across various architectures. Additionally, it is important to note that in BlackVIP, which employs a randomly chosen self-supervised trained encoder in its visual prompt, selecting a suitable encoder architecture that aligns with the backbone architectures of the target pre-trained models is critical, yet challenging in a black-box setting. This suggests that our method may serve as an architecture-independent solution aimed at generally adapting pre-trained visual recognition models.

Methods	RN50	RN101	ViT-B/32	ViT-B/16
ZS	37.5	32.6	45.2	40.8
BAR	26.9	33.5	70.3	77.2
VP	34.7	31.2	71.1	70.9
BlackVIP (RN50)	51.3	50.8	62.9	68.5
BlackVIP (ViT-B/16)	48.1	51.3	67.9	73.1
Ours	54.5	56.1	75.7	80.3

Table 1: Ablation study for backbone architecture of pre-trained models. Classification accuracy on EuroSAT across different backbones of pre-trained CLIP.

REFERENCES

- [1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274* (2022).
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI* 13. Springer, 446–461.
- [3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 105, 10 (2017), 1865–1883.
- [4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [6] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. 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*. IEEE, 178–178.
- [7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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 (2019), 2217–2226.
- [8] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems* 35 (2022), 33716–33727.
- [9] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.
- [10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.
- [11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).
- [12] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.
- [13] Changdae Oh, Hyeji Hwang, Hee-young Lee, Yongtaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. 2023. BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 24224–24235.
- [14] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [17] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. 2020. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*. PMLR, 9614–9624.
- [18] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.
- [19] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.
- [20] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.