Supplementary Materials of "InsVP: Efficient Instance Visual Prompting from Image Itself"

Anonymous Authors

A MORE DETAILS REGARDING HYPER-PARAMETERS

The varying sample sizes among datasets lead us to perform a grid search for optimal learning rate and weight decay. The learning rate and weight decay parameters employed for the ten datasets in the main text are presented in Table 1.

Table 1: The learning rate and weight decay hyper-parameters employed for the ten datasets in the main text. "lr" and "wd" represent the learning rate and weight decay respectively.

	DTD	CUB	Birds	Dogs	Flowers
lr	2e-3	6e-4	9e-4	2e-4	4e-4
wd	1e-1	1e-4	1e-4	1e-4	1e-4

	Food	Cifar100	Cifar10	GTSRB	SVHN
lr	5e-4	3e-3	2e-3	2e-3	5e-3
wd	1e-3	1e-3	1e-3	1e-3	1e-3

B MORE ABLATION STUDIES

B.1 Different Designs of Image-level Instance Prompting.

Table 2: Different designs of image-level instance prompting. The FC, Conv, and Attn represent the fully connected layer, convolutional layer, and attention layer respectively. "Conv-FC" represents the concatenation of Conv and FC, forming a network where Conv and FC are connected sequentially. "FC + Attn" denotes that FC and Attn are two parallel and independent networks.

	CUB	Birds	Cifar100	SVHN
FC	88.5	84.1	90.4	95.3
Conv	88.7	84.2	90.8	95.7
Attn	85.6	81.9	87.9	93.8
Conv-FC	88.9	84.2	90.7	95.5
Attn-FC	86.6	81.6	87.3	93.5
FC + Attn	87.8	83.2	89.5	94.3
FC + Conv	**89.3**	**84.6**	**91.2**	**96.1**

As shown in Table 2, we explore different designs for the instance image prompter using basic network structures such as FC, Conv, and attention layer [1]. It can be observed that using the attention layer resulted in a significant drop in performance. This is because the attention layer has a large number of parameters and requires a substantial amount of data for effective training. On the other hand, using two lightweight networks, FC and Conv, allows us to capture the local patch information and global information of the samples effectively. In future work, we will further explore the network architecture of the instance image prompter to better extract distinctive discriminative information from individual instances.

B.2 Different Designs of Feature-level Instance Prompting.

As shown in Table 3, we conduct comparative experiments with different methods for generating Feature Prompts. Among these methods, the "S-Conv + $\boldsymbol{p}^c$" form achieved the best performance. This is because lightweight convolutional networks have difficulty learning deep features of samples. On the other hand, the common prompt $\boldsymbol{p}^c$ captures common features across all samples, making it relatively easy for the convolutional network to learn the deviations of individual samples from these common features.

Table 3: Different designs of feature-level instance prompting. S-Conv signifies the use of a single convolutional network to generate the feature prompt for all MSA blocks. L-Conv, on the other hand, denotes the utilization of individual convolutional networks for each MSA block to generate the corresponding feature prompt. p^c represents the learnable common prompt introduced in the main text.

	CUB	Birds	Cifar100	SVHN
S-Conv	88.1	83.9	90.1	95.2
L-Conv	88.6	84.1	90.4	95.2
S-Conv + L-Conv	88.4	84.2	90.6	95.4
S-Conv + $\boldsymbol{p}^c$	**89.3**	**84.6**	**91.2**	**96.1**

B.3 Influence of Patch Size in the Patch Prompter.

The patch size in the patch prompter $\mathcal{G}_l$ affects both the model's parameters and the effectiveness of local feature extraction, thus we conducted ablation experiments on it. As shown in Table 5, a patch size of 16 achieved the best performance across all four datasets. This is because when the patch size is further increased, it leads to a rapid increase in the number of parameters in the fully connected layer of the patch prompter. However, as the data in downstream tasks are limited, this can cause the model to overfit, thereby reducing its performance. Conversely, when the patch size is reduced, the number of parameters in the patch prompter decreases and its representational capability is reduced. At the same time, the discriminative information contained in each image patch is reduced, resulting in a decline in the model's performance.

B.4 Different Designs of Weighting Parameters.

fusion weight of image prompts β_I and fusion weight of feature prompt β_F We endeavored to make the fusion weight of image

Table 4: Influence of patch size in the patch prompter $\mathcal{G}_I$.

Patch Size	CUB	Birds	Cifar100	SVHN
2	88.2	82.5	89.7	94.6
4	88.8	83.2	90.5	95.4
8	89.1	83.9	90.9	95.8
16	**89.3**	**84.6**	**91.3**	**96.1**
32	88.9	84.3	91.1	95.9
64	88.4	83.1	90.5	94.4
128	88.1	82.6	89.7	94.2

prompts β_I and the fusion weight of feature prompt β_F learnable parameters for direct optimization. We initialized β_I and β_F to 0.5 and continued learning them. On CUB, the final result was **89.0%**, only marginally lower by 0.3% compared to the fixed parameter. Similar phenomena can also be observed in other datasets. This adaptive weight learning idea can be further explored in future work.

Table 5: Different designs of fusion weight of image prompts β_I and the fusion weight of feature prompt β_F.

	CUB	Birds	Cifar100	SVHN
Fixed	**89.3**	**84.6**	**91.3**	**96.1**
Learnable	89.0	84.2	91.2	95.8

REFERENCES

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