DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting Supplementary Material

Appendix: More Analysis

We provide more analyses of both the design of our model and the training strategies in detail in the section.

Effects of learning rate multipliers. As discussed in Section 4.1, we found that the optimal learning rate for CLIP models and conventional ImageNet pre-trained models are different. Here we further investigate the effects of learning rate multiplier for image encoder and text encoder in Table 1. We see both fixing the text encoder and using a lower learning rate for image encoder is beneficial to train the dense prediction model. Note that we observe a much lower performance ($<$30% mIoU) when directly fine-tuning CLIP models with $1.0\times$ learning rate for the image encoder, which suggests our language guided method can largely stabilize the training process and make the final results less sensitive to the learning rate configuration.

Effects of optimization of the textual contexts. Previous works [1, 3] on transferring CLIP models to downstream classification tasks have clearly shown the importance of adapting the textual contexts for different datasets and tasks. We show the effects of optimizing the textual contexts compared to the original prompting strategy proposed in [2] in Table 2. We see that although the learnable contexts will introduce additional computation during training (gradient computation for the text encoder), this strategy can bring notable improvement over the baseline. Therefore, we choose to add the learnable textual contexts for our models.

Effects of γ. Table 3 shows the effects of γ. We see a learnable γ initialized with small values can improve the final performance.

Table 1. Ablation study of the learning rate multiplier of text encoder and image encoder. We find freezing the text encoder and setting the lr multiplier of image encoder as 0.1 yields the best performance. The configuration used in our final models is highlighted in gray.

	text encoder	image encoder	mIoU (%)
lr multi	0.0	0.1	**43.5**
	0.0	1.0	42.6
	0.1	0.1	42.2

Table 2. Effects of optimization of the textual contexts. We compare the results of using the context optimization [3] with directly constructing textual prompts from human defined template and find learnable textual contexts can bring notable improvements. The configuration used in our final models is highlighted in gray.

Textual Context	mIoU (%)
`a photo of a [CLS].`	42.9
CoOp [3]	**43.5**

Table 3. Ablation study of the residual coefficient γ. The configuration used in our final models is highlighted in gray.

initial value of γ	γ learnable	mIoU (%)
10^{-4}	✗	42.6
10^{-4}	✓	**43.5**
1.0	✓	42.8

References

[1] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 1