Parameter-efficient Fine-tuning in Hyperspherical Space for Open-vocabulary Semantic Segmentation
-Supplementary Material-

Zelin Peng[1], Zhengqin Xu[1], Zhilin Zeng[1], Yu Huang[1], Yaoming Wang[2], and Wei Shen[1(✉)]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Meituan
{zelin.peng, fate311, bernardeschi, yellowfish, wei.shen}@sjtu.edu.cn

This supplementary material is organized into three main sections. First, we present a detailed derivation of the definition introduced in the main paper (Sec. A). Second, we provide additional qualitative results to further demonstrate the contributions of our H-CLIP (Sec. B). Finally, we provide detailed descriptions of the open-vocabulary segmentation datasets in Section C.

A. Derivation of the Definition

In this section, we provide a derivation of definition in the main paper. **Definition 4.1(3-order T-product)** For $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2\times l\times n_3}$, the 3-order T-product $\mathcal{C} \in \mathbb{R}^{n_1\times l\times \times n_3} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) \cdot \text{unfold}(\mathcal{B})), \tag{S-1}$$

where "·" represents standard matrix product.

Definition 4.2(Higher-order T-product) For $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times n_3\cdots\times n_p}$ and $\mathcal{B} \in \mathbb{R}^{n_2\times l\times n_3\times\cdots\times n_p}$, the High-order T-product $\mathcal{C} \in \mathbb{R}^{n_1\times l\times n_3\cdots\times n_p} = \mathcal{A} * \mathcal{B}$ is defined as:

$$\mathcal{C} = \mathcal{A} * \mathcal{B} = \text{fold}(\text{circ}(\mathcal{A}) * \text{unfold}(\mathcal{B})). \tag{S-2}$$

Derivation. According to [3], if $\mathcal{A}$ is $n_1 \times n_2 \times n_3$, $\mathcal{A}$ can be block diagonalized by using Discrete Fourier Transformer (DFT) matrix $\mathbf{F}_{n_3} \in \mathbb{R}^{n_3\times n_3}$ as:

$$(\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \text{circ}(\text{unfold}(\mathcal{A})) \cdot (\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_2}) = \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & & \\ & \ddots & \\ & & \mathbf{D}_{n_3} \end{bmatrix} \in \mathbb{R}^{n_1 n_3\times n_2 n_3}, \tag{S-3}$$

where "⊗" denotes the Kernecker product, "$\mathbf{F}^*_{n_3}$" denotes the conjugate transpose of $\mathbf{F}_{n_3}$, "·" means standard matrix product and $\mathbf{D}$ is a block diagonal matrix. In fact, the i-th block matrix $\mathbf{D}_i$ of $\mathbf{D}$ can be computed by applying DFT of $\mathcal{A}$ along 3-rd dimension. The **3-order T-product** in Eq. (S-1) can be computed as:

$$(\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_1}) \cdot ((\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \text{circ}(\text{unfold}(\mathcal{A})) \cdot (\mathbf{F}^*_{n_3} \otimes \mathbf{I}_{n_2})) \cdot (\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_2}) \cdot \text{unfold}(\mathcal{B}). \tag{S-4}$$

It is readily shown that $(\mathbf{F}_{n_3} \otimes \mathbf{I}_{n_2})$unfold can be computed by applying DFT of $\mathcal{B}$ along 3-rd dimension: the result called $\bar{\mathbf{B}}$. Thus, Eq. (S-4) remains to multiply each block matrix $\mathbf{D}_i$ of $\mathbf{D}$ with each block matrix $\mathbf{B}_i$ of $\bar{\mathbf{B}}$, then take an inverse DFT along the 3-rd dimension of the result. Hence, the **3-order T-product** in Eq. (S-1) can be re-formulated as:

$$\mathcal{C} = \text{DFT}_3^{-1}(\text{DFT}_3(\mathcal{A}) \odot \text{DFT}_3(\mathcal{B})) = \text{DFT}_3^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = \text{DFT}_3^{-1}(\bar{\mathcal{C}}), \tag{S-5}$$

where $\text{DFT}_3(\cdot)$ is DFT along 3-rd dimension and $\text{DFT}_3^{-1}(\cdot)$ is the inverse DFT along 3-rd dimension. In mathematics, the DFT of $\mathcal{A}$ along 3-rd dimension is formulated as:

$$\bar{\mathcal{A}} = \text{DFT}_3(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{F}_{n_3}. \tag{S-6}$$

Similarly, the inverse DFT of $\bar{\mathcal{A}}$ along 3-rd dimension is derived as:

$$\mathcal{A} = \text{DFT}_3^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{F}_{n_3}^{-1}. \tag{S-7}$$

By the detailed theoretical analysis in [4], the DFT has been extended to a general invertible transform S with an invertible transform matrix $\mathbf{S}$. In mathematics, the invertible transform of $\mathcal{A}$ along 3-rd dimension is formulated as:

$$\bar{\mathcal{A}} = \text{S}_3(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_{n_3}. \tag{S-8}$$

Similarly, the inverse transform of $\bar{\mathcal{A}}$ along 3-rd dimension is derived as:

$$\mathcal{A} = \text{S}_3^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}_{n_3}^{-1}. \tag{S-9}$$

Similarly, if $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_p}$, $\mathcal{A}$ can be block diagonalized by using a sequence of DFT matrices $\mathbf{F}_{n_i} \in \mathbb{R}^{n_i \times n_i}, i = 3, 4, \cdot, p$ as:

$$(\mathbf{F}_{n_p} \otimes \mathbf{F}_{n_{p-1}} \otimes \cdots \otimes \mathbf{F}_{n_3} \otimes \mathbf{I}_{n_1}) \cdot \tilde{\mathcal{A}} \cdot (\mathbf{F}_{n_p}^* \otimes \mathbf{F}_{n_{p-1}}^* \otimes \cdots \otimes \mathbf{F}_{n_3}^* \otimes \mathbf{I}_{n_2}) = \mathbf{D}, \tag{S-10}$$

where $\tilde{\mathcal{A}} = \text{circ}(\text{unfold}(\mathcal{A})) \in \mathbb{R}^{n_1 n_3 n_4 \cdots n_p \times n_2 n_3 \cdots n_p}$. Since the matrix $\mathbf{D}$ is block diagonal with $n_3 n_4 \cdots n_p$ blocks each of size $n_1 \times n_2$, the **Higher-order T-product** in Eq. (S-2) can be computed as:

$$(\tilde{\mathbf{F}}^* \otimes \mathbf{I}_{n_1}) \cdot ((\tilde{\mathbf{F}} \otimes \mathbf{I}_{n_1}) \cdot \tilde{\mathcal{A}} \cdot (\tilde{\mathbf{F}}^* \otimes \mathbf{I}_{n_2})) \cdot (\tilde{\mathbf{F}}_{n_3} \otimes \mathbf{I}_{n_2}) \cdot \tilde{\mathcal{B}}, \tag{S-11}$$

where $\tilde{\mathbf{F}} = \mathbf{F}_{n_p} \otimes \mathbf{F}_{n_{p-1}} \otimes \cdots \otimes \mathbf{F}_{n_3}$. Using the DEF, it is straightforward to show that the block diagonal matrix $\mathbf{D}$ in Eq. (S-10) can be obtained by repeated DFTs of $\mathcal{A}$ along each dimension expect for 1-st and 2-nd dimension. Similarly, by using a sequence invertible transform $S_j(\cdot), i = 3, 4, \cdot, p$ with invertible transform matrix $\mathbf{S}_i$, the **Higher-order T-product** in Eq. (S-2) can be re-formulated as:

$$\mathcal{C} = \tilde{S}^{-1}(\tilde{S}(\mathcal{A}) \odot \tilde{S}(\mathcal{B})) = \tilde{S}^{-1}(\bar{\mathcal{A}} \odot \bar{\mathcal{B}}) = \tilde{S}^{-1}(\bar{\mathcal{C}}), \tag{S-12}$$

where $\tilde{S}(\mathcal{A}) = S_p(S_{p-1}(\cdots S_3(\mathcal{A}) \cdots))$, $\bar{\mathcal{C}} = \bar{\mathcal{A}} \odot \bar{\mathcal{B}}$ denotes the frontal-slice-wise product $\bar{\mathcal{C}}(;,;,i) = \bar{\mathcal{A}}(;,;,i) \cdot \bar{\mathcal{B}}(;,;,i), i = 1, 2, \cdots, n_3 n_4 \cdots n_p$ and $\tilde{S}^{-1}(\cdot)$ is the inverse transform of $\tilde{S}(\cdot)$. The inverse transform $\tilde{S}(\cdot)$ is formulated as:

$$\bar{\mathcal{A}} = \tilde{S}(\mathcal{A}) = \mathcal{A} \times_3 \mathbf{S}_3 \times_4 \mathbf{S}_4 \cdots \times_p \mathbf{S}_p, \tag{S-13}$$

and its inverse transform is derived as:

$$\mathcal{A} = \tilde{S}^{-1}(\bar{\mathcal{A}}) = \bar{\mathcal{A}} \times_3 \mathbf{S}_3^{-1} \times_4 \mathbf{S}_4^{-1} \cdots \times_p \mathbf{S}_p^{-1}. \tag{S-14}$$

■

B. Extension Visualization

We present more visualization to illustrate how the misalignment problem impacts segmentation performance, as shown in Figs. S-1 and S-2. These results validate the effectiveness of alignment. In addition, we visualize the training accuracy curve in Fig. S-3, further demonstrating the advantage of the symmetric fine-tuning solution.

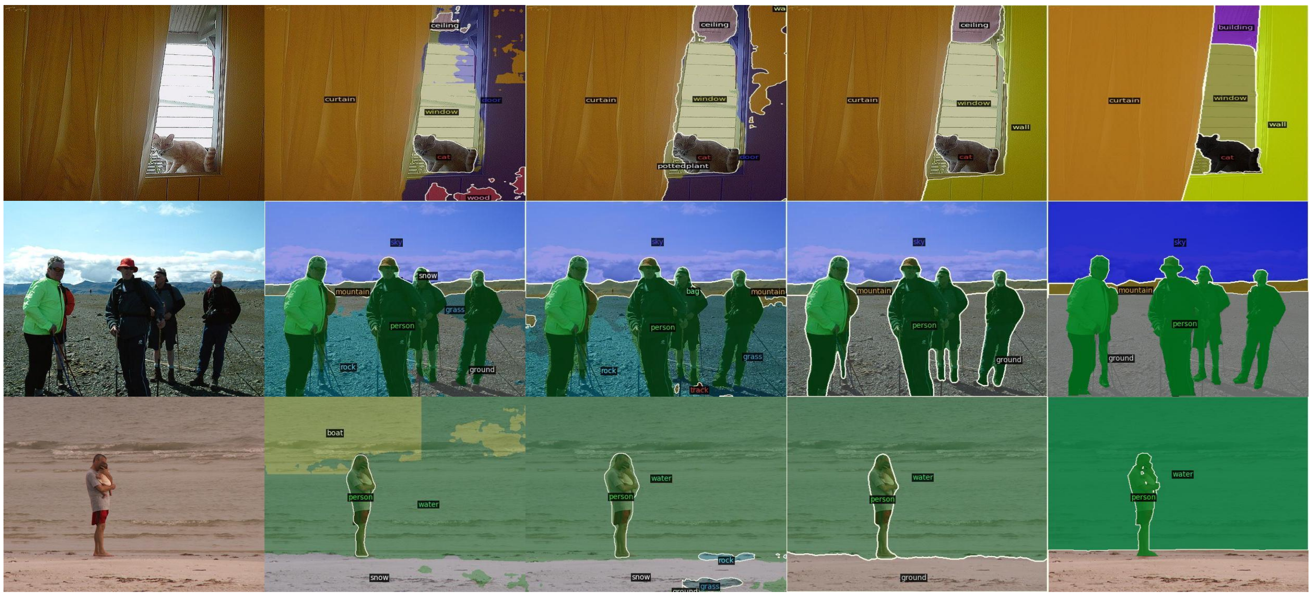

Image Baseline w/o alignment Ours Ground truth

Figure S-1. Comparison of qualitative results on VOC2010 with 59 categories.

Image Baseline w/o alignment Ours Ground truth

Figure S-2. Comparison of qualitative results on ADE20K with 150 categories.

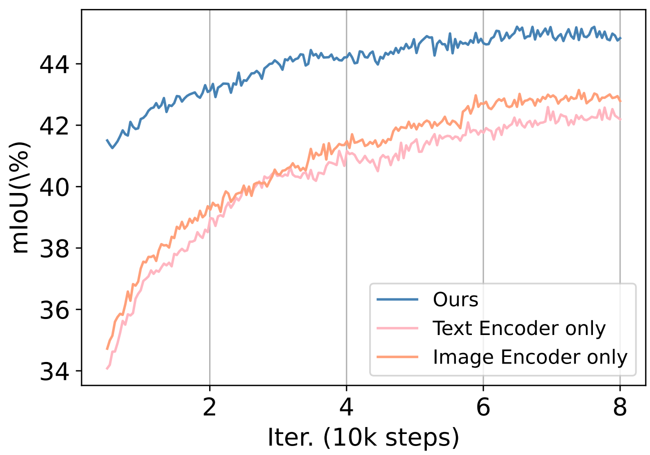

Figure S-3. Training accuracy curves. The comparison is conducted between symmetric fine-tuning (Ours) and asymmetric fine-tuning (text encoder only or image encoder only).

C. Dataset Descriptions

Here, we present detailed descriptions of three datasets we used in open-vocabulary semantic segmentation.

- **ADE20K [6]** is a classical semantic segmentation dataset comprising around 20,000 training images and 2,000 validation images. Besides, it includes two different test sets: `A-150` and `A-847`. The test set `A-150` has 150 common categories, while the test set `A-847` has 847 categories.
- **PASCAL VOC [2]** is a small dataset for semantic segmentation, which includes 1464 training images and 1449 validation images. The dataset contains 20 different foreground categories. We name it as `PAS-20`. In line with [1], we also report a score on $\texttt{PAS-20}^b$, which involves "background" as the 21st category.
- **PASCAL-Context [5]** is upgraded from the original PASCAL VOC dataset. It includes two different test sets: `PC-59` and `PC-459` for evaluation. The test set `PC-59` has 59 categories, while the test set `PC-459` has 459 categories.

References

[1] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2024. 4

[2] 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. 4

[3] Misha E Kilmer and Carla D Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3): 641–658, 2011. 1

[4] Canyi Lu, Xi Peng, and Yunchao Wei. Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5996–6004, 2019. 2

[5] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 4

[6] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 4