Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models Supplementary Material

Yu-Chu Yu[1,†], Chi-Pin Huang[1], Jr-Jen Chen[1], Kai-Po Chang[1], Yung-Hsuan Lai[1], Fu-En Yang[2], and Yu-Chiang Frank Wang[1,2,‡]

[1] National Taiwan University
[2] NVIDIA
[†] r09922104@ntu.edu.tw, [‡] frankwang@nvidia.com

A Evaluation Details

Datasets Statistics. We provide the detailed statistics of 8 fine-grained datasets and the reference dataset (*i.e.*, ImageNet [5]) in Tab. 4. The splits for training, validation and test of each dataset basically follow the setting provided by Zhou *et al*. [13]. Following the setting proposed in ZSCL [12], we sample 100,000 unlabeled images from ImageNet as the reference dataset.

Details of Multiple Training Sequences. We introduce *Multip Training Sequences* evaluation protocol to thoroughly evaluate every method over different training sequences in Sec. 4.3. Here we provide the detailed order of tasks for each sequence in Tab. 5.

B More Implementation Details

Re-Weighted Dual-Teacher Knowledge Distillation Loss. Our proposed Dual-Teacher Knowledge Distillation loss shows the way to select the appropriate teacher model for a reference image according to the dual-teacher discrepancy and selection score η. In practice, there are few reference images with higher dual-teacher discrepancy. To address this potential imbalance problem, we apply a loss re-weighting strategy [4] as a post-processing technique. Specifically, the re-weighted dual-teacher knowledge distillation loss is shown below:

$$\tilde{\mathcal{L}}_{\text{KD}}^{\text{dual}} = \lambda \cdot \sum_{\mathbf{x} \sim \mathcal{X}^{\text{ref}}} \eta(\mathbf{x}) \cdot \mathcal{L}_{\text{KD}}^{k-1} + \sum_{\mathbf{x} \sim \mathcal{X}^{\text{ref}}} (1 - \eta(\mathbf{x})) \cdot \mathcal{L}_{\text{KD}}^{0}, \tag{7}$$

where λ is a hyper-parameter to control the imbalance ratio between the KD loss to the most recent fine-tuned model g_{k-1} and the KD loss to the pre-trained model g_0. Emprically we set $\lambda = 9$ to properly deal with the imbalance issue for every experiment in this work.

Table 4: Detailed statistics for each dataset.

Dataset	Classes	Train	Val	Test
ImageNet [5]	1,000	1.28M	N/A	50,000
Aircraft [8]	100	3,334	3,333	3,333
DTD [3]	47	2,820	1,128	1,692
EuroSAT [6]	10	13,500	5,400	8,100
Flowers-102 [9]	102	4,093	1,633	2,463
Food-101 [1]	101	50,500	20,200	30,300
Oxford-Pets [10]	37	2,944	736	3,669
Stanford-Cars [7]	196	6,509	1,635	8,041
UCF-101 [11]	101	7,639	1,898	3,783

Table 5: The order of tasks for each training sequence.

Sequence	1st Task	2nd Task	3rd Task	4th Task	5th Task	6th Task	7th Task	8th Task
$\mathcal{S}^1$	Aircraft	DTD	EuroSAT	Flowers	Food	Pets	Cars	UCF101
$\mathcal{S}^2$	DTD	EuroSAT	Flowers	Food	Pets	Cars	UCF101	Aircraft
$\mathcal{S}^3$	EuroSAT	Flowers	Food	Pets	Cars	UCF101	Aircraft	DTD
$\mathcal{S}^4$	Flowers	Food	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT
$\mathcal{S}^5$	Food	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers
$\mathcal{S}^6$	Pets	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food
$\mathcal{S}^7$	Cars	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food	Pets
$\mathcal{S}^8$	UCF101	Aircraft	DTD	EuroSAT	Flowers	Food	Pets	Cars

Hyper-Parameters to the η Selection Function. Our proposed η selection function:

$$\eta(\mathbf{x}) = \sigma(\frac{d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x})) - \delta}{\gamma}), \tag{8}$$

involves two hyper-parameters: δ and γ. At a high-level, δ serves as a threshold that determining whether to select more from g_{k-1} or g_0. As the threshold δ increases, more reference data points are likely to be assigned values lower than 0.5, *i.e.*, select KD Loss more from g_0. On the other hand, γ works as a scaling factor to scale the value before applying the sigmoid function. As $\gamma \to 0$, the selection function move towards a *hard selection* mechanism, where the η scores tend to output either 1 or 0, depending on the discrepancy $d(g_{k-1}(\mathbf{x}), g_0(\mathbf{x}))$.

Tab. 6 provides a sensitivity analysis for hyper-parameters δ and γ. In general, the performance shows no significant difference when $\delta = 0.1$ or 0.2, hinting that it is stable enough for a proper range. By default, we select $\delta = 0.2$ and $\gamma = 1/6$ across all experiments in this work.

C Different Choices of Reference Datasets

Our *Selective Dual-Teacher Knowledge Transfer* framework leverages an unlabeled reference dataset, following the settings in [12]. As mentioned in the limi-

Table 6: Sensitivity analysis on $\mathcal{S}^1$ to the hyper-parameters δ and γ in the η selection function. We highlight the results of our default setting across all experiments in the main paper in light red.

δ	γ	Forgetting (↓)	Degradation (↓)	Avg. Accuracy (↑)
0.1	1/3	1.72	1.58	84.42
	1/6	1.68	1.57	84.43
	1/9	**1.65**	1.58	84.47
0.2	1/3	1.67	1.60	84.46
	1/6	1.70	**1.55**	**84.48**
	1/9	1.82	1.86	84.31
0.3	1/3	1.81	1.52	84.23
	1/6	2.13	1.99	84.03
	1/9	2.45	1.99	83.93

Table 7: The performance of different reference datasets with varying size. The default setting for all experiments is marked in light red.

Ref. Dataset	Size	Forgetting (↓)	Degradation (↓)	Avg. Accuracy (↑)
ImageNet	10k	1.92	2.12	84.18
	100k	1.70	1.55	84.48
	200k	1.65	**1.11**	84.80
Conceptual Captions 12M	10k	2.28	2.17	83.84
	100k	**1.50**	1.88	84.48
	200k	1.60	1.25	**84.99**

tation, the composition and the diversity of the images in the reference dataset might greatly affect the final performance. To examine the effect, we conduct ablation studies using different reference datasets (e.g., ConceptualCaptioning 12M [2]) and exploring the impact of varying the size of the reference dataset. Tab. 7 shows the performance of different reference datasets with varying size. While increasing the size of the reference dataset typically enhances performance, empirically there are no significant differences when the size exceeds 100k. By default, we use ImageNet with 100k images as our reference dataset, which also aligns with the same settings in [12].

D Experiments Details

Detailed Explanation to the Visualization of Reference Images with Large η Scores. To illustrate the reference images with the highest η scores, we train our model on the first sequence $\mathcal{S}^1$ (the detailed task orders are shown in Tab. 5). For each stage $k \geq 2$, we calculate the η scores for each reference image using **only** the original pre-trained model g_0 and the most recent fine-tuned model g_{k-1} according to Eq. (2). Then, we select the Top-25 images with the highest η scores. Given that the visual concepts in some datasets are challenging

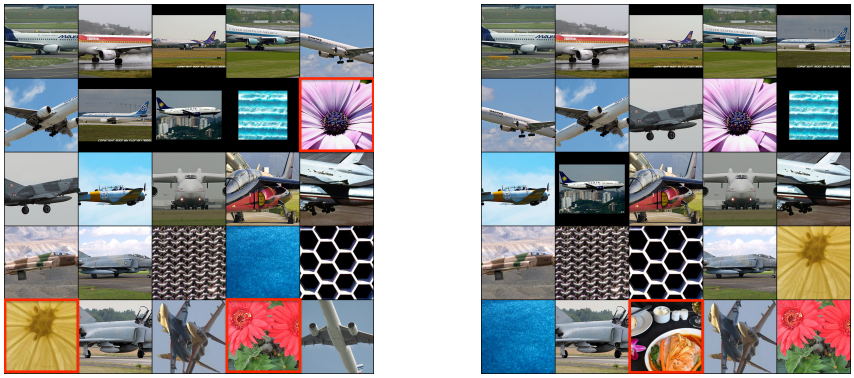

Fig. 7: Example images selected from the reference dataset with large η scores. **Left**: Top-25 reference images selected *after* fine-tuning on the Flowers Dataset. **Right**: Top-25 reference images selected *after* fine-tuning on the Food Dataset.

to depict (*e.g.*, EuroSAT, UCF101), we focus our visualizations on datasets with more concrete concepts, such as Flowers and Food, as visualized in Fig. 7.

Detailed results for Catastrophic Forgetting and Zero-Shot Degradation. In Fig. 4 and Fig. 5, we present examples of the assessment of catastrophic forgetting for the first task and evaluation of zero-shot degradation for the last task, respectively. Here we plot the impact of catastrophic forgetting on the first task and the impact of zero-shot degradation on the last task across each sequence in Fig. 8 and Fig. 9. For catastrophic forgetting, our method clearly outperform other methods by stably preserving the performance on the previously fine-tuned task (1st task in this case). Regarding the issue of zero-shot degradation, our method effectively maintains the original zero-shot capabilities in most scenarios, highlighting our success in preserving both pre-trained and previously fine-tuned knowledge across diverse datasets and various sequences.

E The Training Algorithm of Our Proposed Framework

As discussed in Sec. 3.3, we provide the detailed training algorithm of our *Selective Dual-Teacher Knowledge Transfer* framework in Algorithm 1.

Algorithm 1 Selective Dual-Teacher Knowledge Transfer

Input: A pre-trained VLM g_0, hyper-parameters $\delta, \gamma, \lambda_{\text{dual}}$.
Data: A sequence of training tasks $\mathcal{S} = (\mathcal{T}^1, \cdots, \mathcal{T}^K)$ and a reference dataset $\mathcal{X}^{\text{ref}}$.
Output: The final fine-tuned model g_K.

for k in $1 : K$ **do**
  Freeze g_0 as the pre-trained knowledge teacher.
  Freeze g_{k-1} as the previously fine-tuned knowledge teacher.
  Initialize the current model g_k by g_{k-1}.
  for e in E **do**
    while not traverse over all current data $\mathcal{T}^k$ **do**
      Sample a batch of current data B^k.
      Sample a batch of ref data B^{ref}.
      Calculate $\mathcal{L}_{\text{CE}}$ with the current data B^k.
      Calculate Eq. (3) with g_0, g_{k-1}, and B^{ref}.
      Update g_k with loss function Eq. (4).
    end while
  end for
end for

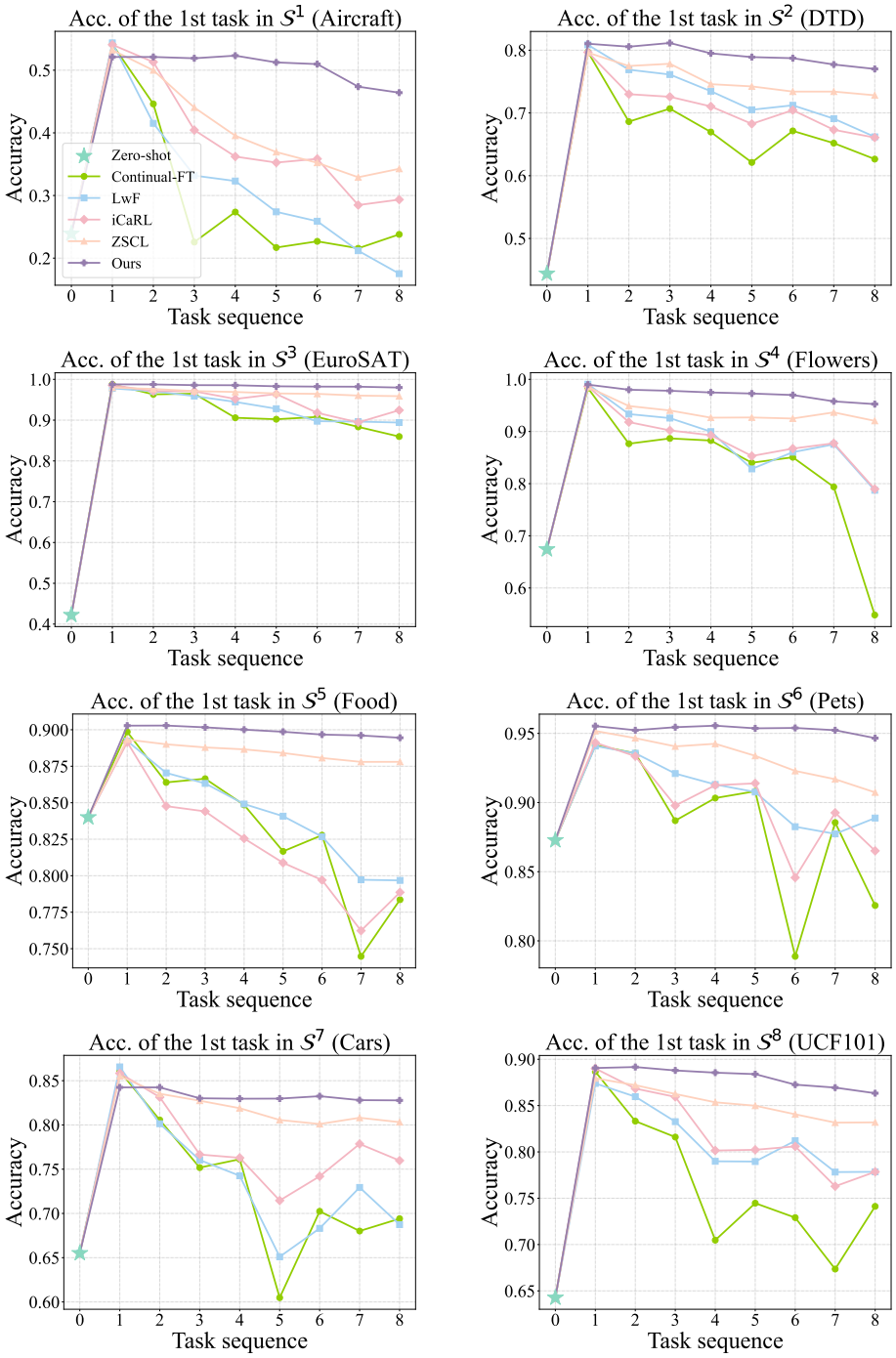

Fig. 8: Assessment of catastrophic forgetting with the first task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method is able to maintain their accuracies at the end of learning sequence.

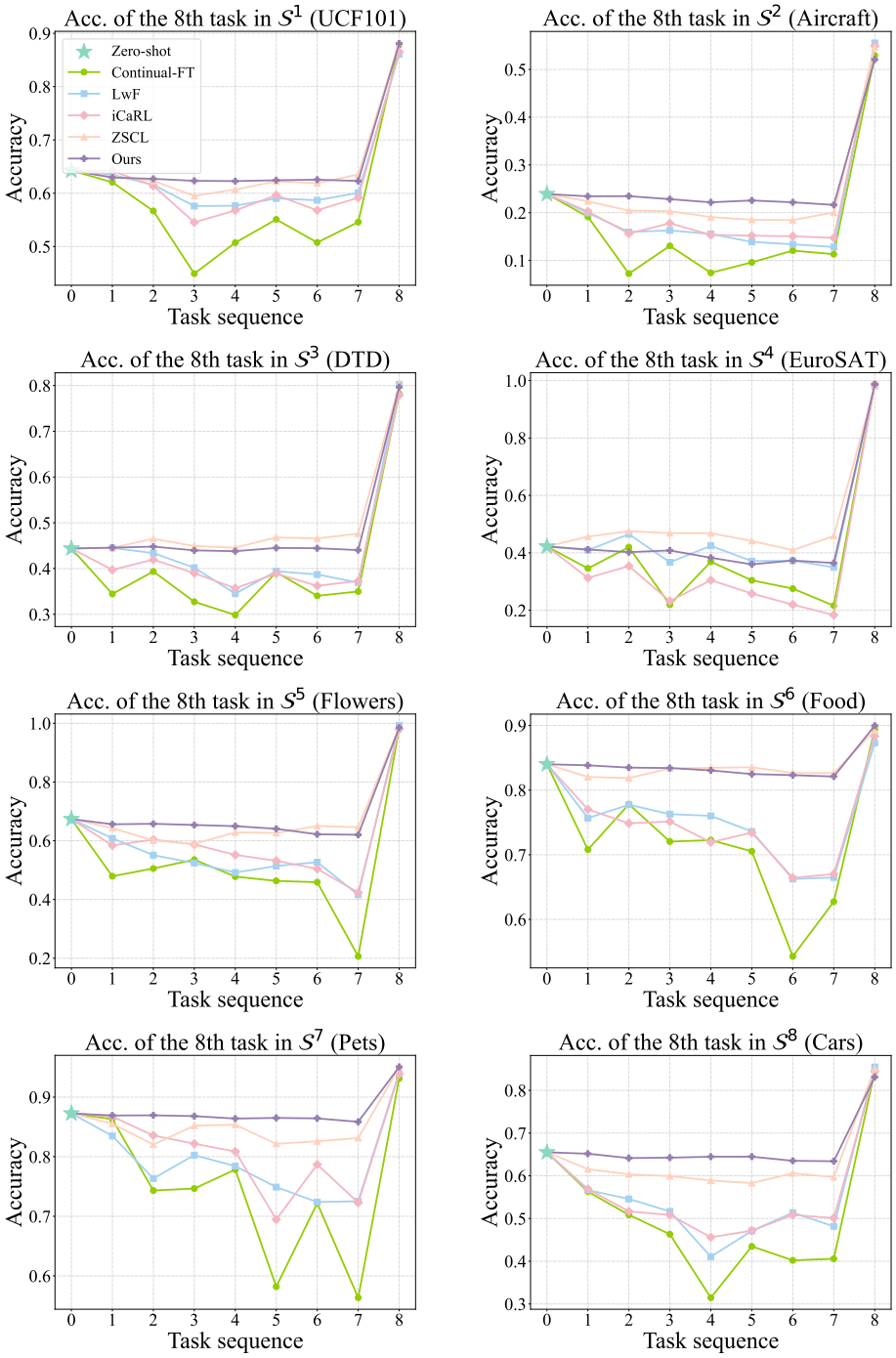

Fig. 9: Assessment of zero-shot degradation with the last task in the continual learning sequence (i.e., the horizontal axis). It can be seen that our method shows satisfactory accuracies before finetuning on the last task

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: 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. pp. 446–461. Springer (2014)
2. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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), 2217–2226 (2019)
7. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
8. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
9. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
10. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
11. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
12. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023)
13. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)