Learning from the Web: Language Drives Weakly-Supervised Incremental Learning for Semantic Segmentation
Supplementary Material

Chang Liu[1], Giulia Rizzoli[2], Pietro Zanuttigh[2], Fu Li[1], and Yi Niu[1]*

[1] School of Artificial Intelligence, Xidian University, China
[2] Department of Information Engineering, University of Padova, Italy

This document contains the supplementary material for the paper *Learning from the Web: Language Drives Weakly-Supervised Incremental Learning for Semantic Segmentation*. Firstly, we provide a detailed description of the knowledge distillation losses of the localizer and segmentation model. Then, to present further validation, we showcase 1) an upper bound of the proposed method that directly uses PASCAL-VOC as the replay source and 2) the per-step results on the VOC dataset obtained in the most challenging setting (10-1), which involves the highest number of incremental steps. Besides, we also explore the robustness of the proposed method to the choice of different VLMs. Finally, we motivate our framework by showing ablation studies, quantitative and qualitative support for each component of the method, including the Fourier discriminator, caption pseudo-labeling, caption downloading, and caption filtering.

1 Knowledge Distillation Losses

The localizer L, introduced in [3], is used to provide a pseudo-supervision for the main segmentation model. It shares the same encoder E with main model and predicts a score $\mathbf{y}_L$ for all the classes $|\mathcal{Y}|$:

$$\mathbf{y}_L = (E \circ L) \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{Y}|} \tag{1}$$

The training objective of the localizer has a dual role: 1) learning new classes from image-level labels and 2) refining old classes using the previous segmentation model's output. The former follows the classification task, in which the class score is normally calculated with global average pooling (GAP). However, using GAP tends to encourage all the pixels to identify with the target classes, which weakens diversity and is thus not suitable for the segmentation task. Based on the above concern, Araslanov et al. [1] proposed Global Weighted Pooling (GWP) and focal penalty to calculate a more precise classification score. While effective for new classes, this approach still struggles with localizing previous classes. To this extent, WILSON [3] introduced a knowledge distillation-based localization

* Corresponding author

prior loss (KDL) that leverages the previous segmentation model's output as pseudo-supervision for old classes:

$$\mathcal{L}_{KDL}(\mathbf{z}, \mathbf{y}_D^{(t-1)}) = -\frac{1}{|\mathcal{Y}^{t-1}||\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{Y}^{t-1}} y_{(i)D}^{c(t-1)} log(\sigma(z_i^c)) + \\ + (1 - y_{(i)D}^{c(t-1)}) log(1 - \sigma(z_i^c)) \tag{2}$$

where σ denotes the logistic function applied to $\mathbf{z}$, which represents $\mathbf{y}_L$ prior to the softmax operation, and $\mathbf{y}_D^{(t-1)}$ is the output of the segmentation model from step $t-1$. The KDL loss provides pixel-wise supervision to the localizer, enhancing the precision of the localizer's class activation map. Similarly, a distillation loss $\mathcal{L}_{KDE}$ is applied to the segmentation model: it computes the mean-squared error between the features extracted by the encoder of the current step E^t and the one of the previous step E^{t-1}:

$$\mathcal{L}_{KDE} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|e_i^t - e_i^{t-1}\| \tag{3}$$

where e is the feature vector from the encoder E.

2 Additional Results

2.1 VOC as Rehearsal data

We present in Table 1 the results obtained by utilizing VOC directly for replay, which serves as an upper bound for our approach. Note how our approach closely approaches the bound, particularly in the 15-5 setting, demonstrating the effectiveness of utilizing WEB data. Furthermore, when employing 50 VOC images — equivalent to the number used by the FMWISS [7] competitor — our approach yields slightly better or similar results. Specifically, we observe a slight decrease of 0.3 mIoU points in the 10-10 setting and an improvement of 0.8 in the 15-5 setting compared to FMWISS.

Table 1: Results using VOC as replay source on PASCAL-VOC overlapped setup.

				10-10			**15-5**		
Method	**Train**	**Memory**	**M**	1-10	11-20	all	1-15	16-20	all
FMWISS [7]	VOC	VOC	50	**73.8**	62.3	<u>69.1</u>	**78.4**	54.5	73.3
Ours	VOC	WEB	100	<u>73.6</u>	55.5	65.7	<u>78.2</u>	54.9	73.3
Ours	VOC	VOC	50	67.4	**68.1**	68.8	76.5	<u>63.4</u>	<u>74.1</u>
Ours	VOC	VOC	100	71.9	<u>66.2</u>	**70.1**	76.5	**65.5**	**74.6**

2.2 Class-wise Per-step mIoU

Fig. 1 illustrates the per-step performance of our approach (solid lines) and compares them with WILSON [3] (dashed lines), highlighting the consistent improvement of our method over WILSON across the various steps of the learning sequence. We outperform the competitor by a large margin in most classes (including the initial set of classes learned in the first step) even if a very few challenging ones persist, such as dining table and potted plant.

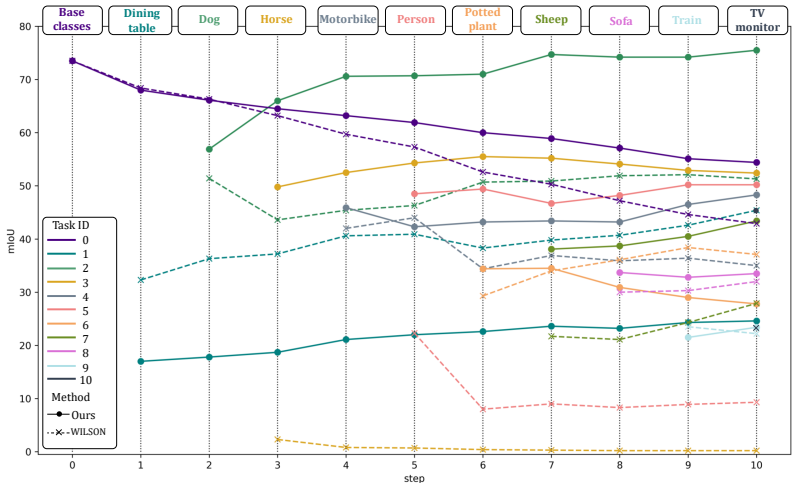

Fig. 1: Per-task and per-step mIoU for the 10-1 VOC multi-step *overlap* incremental setting (WEB+WEB).

2.3 Evaluation with a different VLM

While the results in the paper were obtained using the OpenFlamingo [2] model, our approach demonstrates versatility to different Vision-Language Models (VLMs). In Tab. 2 we showcase the results of replacing OpenFlamingo with BLIP [5]. The performances are slightly better in the 10-10 setting and slightly worse on average in 15-5. Notably, these performance differentials fall within a narrow range of [0.2, 0.5] mIoU points. These minimal variations confirm that our proposed method is not overfitted on a particular VLM configuration. Rather, it shows the robustness and generalizability in combination with different models.

Table 2: Comparison with BLIP VLM in the PASCAL-VOC overlapped setup.

Method	Train	Memory	M	10-10			15-5		
				1-10	10-20	all	1-15	16-20	all
OpenFlamingo	VOC	WEB	100	**73.6**	55.5	65.7	**78.2**	**54.9**	**73.3**
BLIP	VOC	WEB	100	73.3	**56.3**	**65.9**	78.1	54.1	73.0
OpenFlamingo	WEB	-	-	**73.8**	53.9	65.0	**74.4**	45.9	**68.4**
BLIP	WEB	-	-	72.2	**56.1**	**65.3**	73.5	**46.4**	67.9
OpenFlamingo	WEB	WEB	100	**73.7**	54.5	65.3	**78.3**	47.8	71.7
BLIP	WEB	WEB	100	73.0	**55.8**	**65.6**	77.1	**51.9**	**71.8**

3 Framework motivation

In this section, we aim to support each component of our method by presenting some additional quantitative and qualitative results.

3.1 Learning new Knowledge from Web

Fourier Domain-based Discriminator To extract images with statistics resembling the original ones, we have trained a discriminator network, helping us to select web images for the new classes that closely resemble those from the original dataset. We trained the discriminator until it reached 80% accuracy on VOC data to ensure a proper behaviour of this model. To validate the efficacy of training the discriminator in the Fourier domain, we compare in Tab. 3 the performance of our approach with the baseline strategy in the pixel domain, i.e., substituting Eq. 6 of the main paper with $(p_{ds}, p_{web}) = M_D(\mathbf{x})$. In the frequency domain, the general style information of the image is primarily contained in lower frequencies, while specific content is typically associated with higher frequencies [6]. This separation makes distinguishing dataset-wide properties from image-specific content easier than using approaches in the pixel domain or exploiting the Fourier phase (see Tab. 3). Using the amplitude values of the Fourier Transform in place of the original sample, values lead to a mIoU improvement in learning new classes of 3.3% on 10-10 and 2.3% on 15-5 and of 1.2% and 0.9% in the average value. Moreover, this strategy demonstrates good generalization, with accuracy on new (unseen) classes during incremental steps reaching $76 - 77\%$, quite close to the 80% achieved on training classes. In contrast, with a pixel-domain discriminator we only achieved $64 - 66\%$ accuracy.

Caption Labeling Although the images for the new classes are queried by the class name (Eq. 4), the downloaded image might also contain classes that have already been learned or are currently being learned. The latter case is particularly evident in the COCO-to-VOC scenario, where 20 classes are being learned simultaneously. In Fig. 2 we provide examples demonstrating how, by analyzing captions and cross-referencing words with a reference dictionary (see Tab. 4), we can derive multi-class labels.

Table 3: Sample selection strategy in the spatial and Fourier domain: mIoU in the PASCAL-VOC overlapped setup.

	10-10			15-5		
Domain	1-10	11-20	all	1-15	16-20	all
Pixel	**72.6**	48.6	61.9	73.2	44.0	67.1
Fourier-Phase	72.4	50.1	62.4	**73.9**	43.8	67.6
Fourier-Ampl.	72.0	**51.9**	**63.1**	73.7	**46.3**	**68.0**

Fig. 2: Image-level labels generated from captions for COCO-to-VOC incremental step classes. For each sample we show (from top to bottom) the queried class name, a thumbnail of the image, the generated caption and the final image-level label.

Table 4: Set of words $\mathcal{W}$ corresponding to each class c in PASCAL-VOC.

class	synonyms	class	synonyms
airplane	plane, jetliner	dining table	-
bicycle	bike	dog	-
bird	parrot, duck, flamingo, swan, seagull, chicken	horse	-
boat	ship	motorbike	motorcycle
bottle	-	person	man, men, woman, women, people, baby
bus	-	potted plant	pot of plant, pot of flower
car	-	sheep	-
cat	-	sofa	couch
chair	-	train	train car
cow	-	tv	television, monitor, tv monitor

3.2 Reharsal strategies

Caption-based Querying. In Fig. 3 we show some examples of downloaded images obtained through class-name (i.e., Eq. 4) and through the caption (i.e., Eq. 8). Comparing them visually reveals that the caption method retrieves images more similar to the PASCAL-VOC ones without needing the storage of sensitive data. Specifically, in the first example the caption preserves related objects from the original scene (i.e., the lamp and the desk). In the subsequent two instances, it maintains the specific plane and car model. Finally, in the last two cases, it keeps the original background (i.e., the grass and the water). Notably, the background class - including not only the *stuff* classes but all the objects that are not part of the training - plays an important role in causing the distribution shift as underlined in previous CILSS research works [4].

Caption-based Filtering. Although our web retrieval scheme is efficient, there is no definite assurance that the downloaded images actually contain the objects referenced in the respective captions q'. This discrepancy can be seen in Fig. 4, where the leftmost images show the original PASCAL-VOC image for which the captions were generated, and each upper row contains examples of downloaded images for which the re-generated captions q'' do not match the original one. Our method verifies this consistency by not only retaining the main objects (such as the dog, the sofa, and the person) but also capturing contextual details in some instances, for instance, the living room and the mountain. More examples are shown in Fig. 5 that compares PASCAL-VOC samples with similar ones retrieved from the web. Additionally, the ablation on parameter T of the caption-filtering is in Tab. 5. Results are stable for T in $[0.5, 0.7]$, with $T = 0.6$ leading to the best tradeoff between old and new classes. Furthermore, we conducted an ablation study on the number of selected words from the caption at step 2 of the filtering strategy. Our findings indicate that considering only 2 words proved to be slightly preferable. This outcome may be attributed to the characteristics of the VOC

Fig. 3: Comparison between caption and class-based web query strategy. (Left) Pascal sample and its related class and caption. (Upper right) Web samples downloaded with class name. (Lower right) Web samples downloaded with caption.

dataset, where less than 1/10th of the images contain more than two classes. The parameter can be tuned depending on the target distribution data. Nevertheless, when the distribution is unknown, all the nouns can be kept, leading to similar results as shown in Tab. 6.

Table 5: Ablation on threshold T in caption-based filtering: mIoU in the PASCAL-VOC overlapped setup.

	10-10			**15-5**		
T	1-10	10-20	all	1-15	15-20	all
0.0	72.2	55.1	64.9	77.6	53.9	72.6
0.5	72.5	**55.8**	65.3	77.8	53.9	72.8
Ours (0.6)	**73.6**	55.5	**65.7**	**78.2**	**54.9**	**73.3**
0.7	73.4	55.5	65.6	78.1	53.6	72.9

Table 6: Ablation on the number of nouns N in caption-based filtering: mIoU in the PASCAL-VOC overlapped setup.

	10-10			**15-5**		
N	1-10	10-20	all	1-15	15-20	all
1	73.0	55.8	65.6	77.9	52.9	72.6
Ours (2)	**73.6**	55.5	65.7	**78.2**	**54.9**	**73.3**
3	73.1	55.9	65.7	78.1	53.8	73.0
ALL	72.5	**56.8**	**65.8**	77.8	54.0	72.8

Fig. 4: Examples of selected and discarded images by regenerating caption for the downloaded web samples. (Left) Dataset image and its caption; (Upper right) Discarded samples and corresponding captions. (Lower right) Selected samples and corresponding captions.

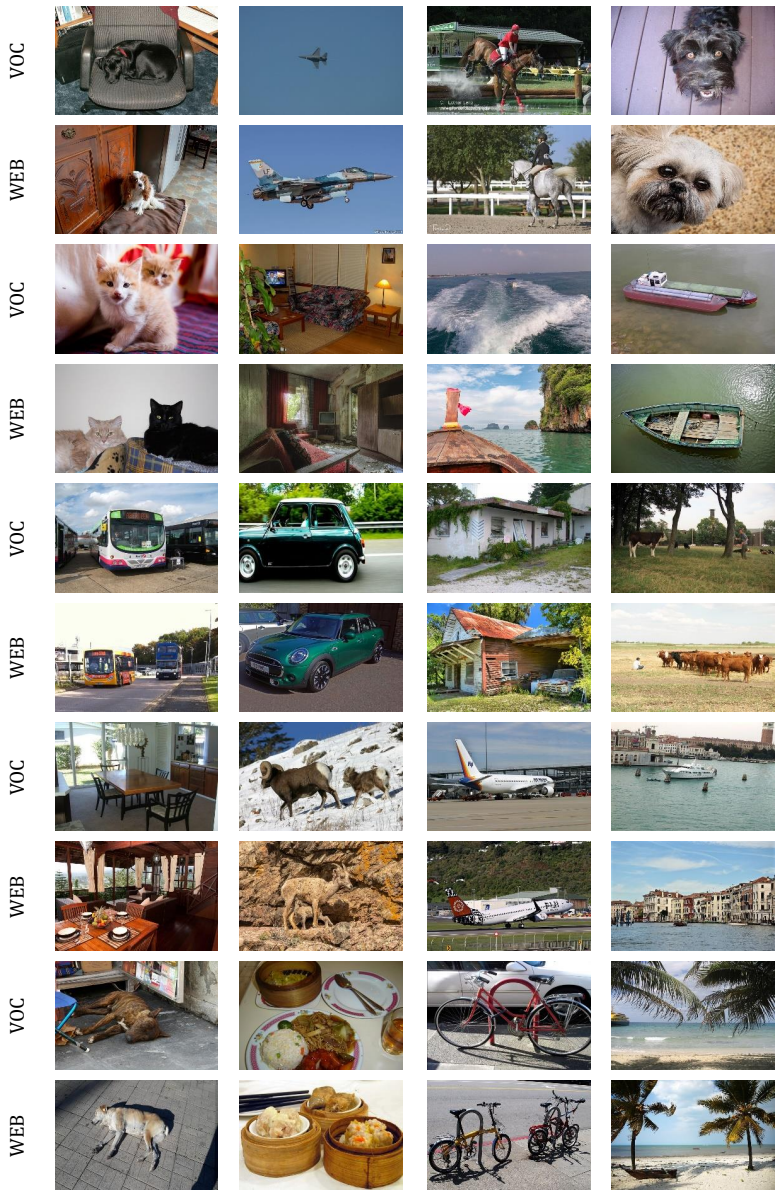

Fig. 5: Replay samples selected with caption model compared with PASCAL-VOC ones.

References

1. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4253–4262 (2020)
2. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
3. Cermelli, F., Fontanel, D., Tavera, A., Ciccone, M., Caputo, B.: Incremental learning in semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4371–4381 (2022)
4. Cermelli, F., Mancini, M., Bulò, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: CVPR (2020)
5. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
6. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4085–4095 (2020)
7. Yu, C., Zhou, Q., Li, J., Yuan, J., Wang, Z., Wang, F.: Foundation model drives weakly incremental learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23685–23694 (2023)