MS3: A MULTIMODAL SUPERVISED PRETRAINED MODEL FOR SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to the limited labeled data, current segmentation models are usually transferred from ImageNet pretrained models. This pipeline introduces task gaps, where the pretraining is based on global image-level recognition while the downstream is focused on local pixel level prediction. In this paper, we aim at mitigating this task gap and building a segmentation-oriented pretrained model, in this way different downstream segmentation tasks can be better and easily adapted. Towards this goal, we combine off-the-shelf annotations from diverse segmentation datasets and make use of both visual and language supervision for jointly training. The highlight is that the two kinds of supervision are complementary and can be boosted to better model the class relation from diverse datasets. The proposed learning framework, termed as **MS3** (short for **M**ultimodal **S**upervision for **S**emantic **S**egmentation), not only adjusts and improves the quality of language embeddings to fit the segmentation scene, but also generates momentum-updated visual embeddings for each category to facilitate better visual representation modeling. Besides, considering that the original one-by-one pixel-embedding pairing may cause similar classes from other datasets to be incorrectly pulled away, we further extend the original loss with multi-label mapping via cross-modal information exchange to better model the class relations. Experiments conducted on several benchmarks demonstrate that MS3 consistently outperforms the ImageNet pretrained models by a considerable margin under standard fine-tuning, as well as fitting some rapid deployment scenarios, *e.g.*, frozen-backbone fine-tuning or zero shot predicting.

1 INTRODUCTION

As a fundamental task in computer vision, semantic segmentation has witnessed great success over the past decades. However, training segmentation models usually require a large number of pixel level annotations, which is time-consuming and difficult to accumulate. Therefore, standard segmentation frameworks usually rely on transfer learning with models fine-tuned on pretrained weights, *e.g.,* training on a large-scale ImageNet dataset (Russakovsky et al., 2015) in a supervised (He et al., 2016; Dosovitskiy et al., 2020; Liu et al., 2021) or unsupervised (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Grill et al., 2020) manner. While this pipeline can *de facto* bring performance gain and avoid overfitting to some extent, it suffers from *task gap* that the pretraining is based on global discrimiation while the downstream task is focused on local pixel level prediction. Such discrepancy limits the transferring ability of the pretrained model.

To mitigate the task gap, it is appealing to directly build a segmentation-based pretrained model. Some recent works introduce pixel-level contrastive learning (Chen et al., 2020a; He et al., 2020; Wang et al., 2021b) for semantic segmentation pretraining (Xie et al., 2020; Wang et al., 2021c; Zhao et al., 2020), which is achieved by pulling close the pixel level embeddings with the same semantic and pushing apart pixels with different semantics. However, these works are constrained within a single dataset, and due to the lack of abundant pixel-level annotations, their generalization ability are still limited. An alternative is to utilize the available annotations off-the-shelf from diverse segmentation datasets for jointly training. However, Lambert et al. (2020) demonstrates that simply training on the merged datasets may not drive the desired performance, as different datasets suffer inconsistent class definitions. Lambert et al. (2020) further proposes to re-label the merged dataset with a manually created unified label-set, but it requires additional labor costs and is error-prone.

Table 1: Analysis of similarity for language embeddings between two closely-connected class pairs and the mean similarity between all the embedding pairs. $\mathbb{M}_I$ represents the proposed module for language embedding improvement.

Method	Cosine Similarity		
	'cat' vs. *'dog'*	*'horse'* vs. *'cow'*	mean
CLIP	0.92	0.92	0.77
CLIP + $\mathbb{M}_I$	0.29	0.17	0.11

In this paper, we follow the idea of unifying multiple datasets for segmentation pretraining and hope to construct a segmentation-oriented model that benefits from it, diverse downstream tasks can be efficiently adapted. The core challenge is to associate the relevant class that defined in different datasets. To this end, we intuitively extend segmentation pretraining with the language embedding supervision. Since relevant classes can be automatically discovered and better modeled via the text expression, and thus transferred to the visual branch can reduce conflicts. Though promising, simply borrowing the text encoder such as CLIP (Radford et al., 2021), which like most previous language-driven works do (Li et al., 2022; Yin et al., 2022), would bring limited benefits under the pretraining-finetuning pipeline. As shown in Tab. 1, we find that some closely-connected class pairs (*e.g.*, *'cat'* vs. *'dog'*, *'horse'* vs. *'cow'*) exhibit very high similarity when directly using CLIP as embeddings, which makes them indistinguishable. This is possibly due to the restriction of the human-designed templates that overwhelm the discrepancy of different classes, as well as the mismatched pretraining objectives, where the CLIP is based on image-level pairing which may aggregate redundant background information, while the target is focused on pixel level comparison.

To address this issue, we propose a multi-modal supervision scheme to better meet the pre-training scenarios. We first adapt the vanilla language-driven scheme and improve the quality of CLIP language embedding to fit the segmentation scene. Different from the original CLIP that conducts image-sentence pairing and uses human-designed templates like "a photo of a [CLS]." as text prompts, we introduce learnable textual contexts to the prompts and conduct pixel-language pairing on segmentation datasets to obtain higher quality language embeddings. Meanwhile, considering that some language-indistinguishable categories may be more visually discriminative, we additionally generate a momentum-updated visual embedding for each category as complements to the language embeddings. With the combined supervision of vision and language, the relevant classes among different datasets can be well considered. Besides, we also notice that the original pixel-embedding pairing loss cannot make full use of the refined relationship due to its one-by-one hard pairing property, as similar class embeddings in other datasets will be incorrectly pushed away. Based on the multimodal embeddings, we further design a novel cross-modal information exchange module and generate multi-label mapping to alleviate this problem.

The above proposed framework, termed as **MS3** (short for **M**ultimodal **S**upervision for **S**emantic **S**egmentation), can serve as a pretraining model for standard downstream fine-tuning, and achieves noticeable performance gains compared to ImageNet pretrained models. Furthermore, since we explore and exploit the relations between classes during pre-training, MS3 can also achieve satisfactory performance under frozen-backbone fine-tuning or zero shot learning scenarios for rapid deployment. Experiments conducted on several benchmarks demonstrate the effectiveness of MS3.

In a nutshell, this paper makes the following contributions:

- We propose a novel multi-dataset pretraining framework for semantic segmentation with the help of multi-modal supervision, in which modules are designed to both improve the quality of language embeddings and bring in additional visual embedding supervision for better pixel-level feature learning.
- A novel cross-modal information exchange module that provides multi-label extension is also proposed to alleviate the performance bottleneck caused by the one-by-one pairing loss under the multi-dataset situation.
- Our framework consistently outperforms the pretrained models over ImageNet by a considerable margin under standard fine-tuning. It can also achieve satisfactory performance under rapid deployment scenarios. We hope our simple but effective framework would shed light on task adaptive pretraining research.

2 RELATED WORK

Segmentation-oriented pretraining. Some recent works extend global-level contrastive learning to the pixel level for segmentation-oriented pretraining. Among them, Zhao et al. (2020) pulls close the pixel level embeddings with the same class labels and pushes apart embeddings with different labels. However, the pretraining is only conducted on a single segmentation dataset, and the transferability is limited by the pre-training scale. DenseCL (Wang et al., 2021c) and PixPro (Xie et al., 2020) respectively construct positive sample pairs using pixel similarity and pixel distance as pseudo-labels on a large-scale ImageNet dataset, while the reliability of pseudo-labels restricts the performance.

Multi-dataset semantic segmentation. For multi-dataset segmentation, Wang et al. (2021a) uses dataset-specific classifiers to alleviate the influence of label differences, while they ignore the label relationships between different datasets. Some works further conduct manual label unification to deal with class differences. Liang et al. (2018) builds a semantic concept hierarchy by combining labels from four datasets and explicitly incorporates the hierarchy into network construction. MSeg (Lambert et al., 2020) unifies the taxonomies of seven semantic segmentation datasets by manually identifying a new class label for every segmentation. To avoid the tedious manual operations, Yin et al. (2022) generates text embedding for each class and uses them as supervision, and the relationships between categories from different datasets are implicitly modeled in the embedding space. Shi et al. (2021) changes the original pixel classification objective and designs a supervised contrastive-like loss and some cross-dataset schemes for multi-dataset learning. Kim et al. (2022) generates multi-label mapping and purposes a novel Class-relational BCE loss to reduce the gradient conflict issue caused by label differences. Our work also does not require label unification, but differently, we design a multimodal scheme and explicitly exploit both cross-dataset and cross-modal category relationships.

Language-driven recognition. CLIP (Radford et al., 2021) is a milestone work for language-driven recognition, which demonstrates that classic recognition tasks that are not commonly associated with language can strongly benefit from language assistance. CLIP jointly trains an image encoder and a text encoder to predict the correct image-text pairings and synthesize extremely robust models for zero-shot image classification. Recent works have also extended this basic paradigm to perform flexible object detection (Gu et al., 2021; Gao et al., 2021; Zhong et al., 2021) and segmentation (Rao et al., 2022). Among them, some language-driven segmentation methods (Li et al., 2022; Yin et al., 2022) directly leverage high-capacity language models to embed the descriptive input labels and train the image encoder to maximize the correlation between the image pixel embeddings and their corresponding text embeddings, which also enables a flexible segmentation. Our work is inspired by them, but we bring the modeling capacity of multimodal embeddings into the multi-dataset pretraining scenario and further conduct some targeted designs.

3 METHOD

3.1 OVERVIEW

In this section, we elaborate on the multi-modal supervised pretraining pipeline for the semantic segmentation model. As illustrated in Fig. 1, we unify multiple segmentation datasets for jointly training. For vision branch supervision, we choose off-the-shelf annotation for each pixel, and conduct pixel-embedding pairing with visual embeddings generated for each category. While for language supervision, considering the representation gap between the general CLIP model and the merged class labels from multiple segmentation datasets, we first adapt and improve the original CLIP language embeddings to the target segmentation datasets, and follow the same pixel-embedding pairing as vision supervision but using the improved language embedding as the objective for each category. In addition, we further conduct multi-label loss extension by cross-modal information exchange to better model the inter-class relationship.

Specifically, given multiple segmentation datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ as well as their corresponding label space $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_N$, we first unify them into a large dataset $\mathcal{D}$ with a combined label space $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2 \ldots \cup \mathcal{Y}_N$,where N is the number of datasets. Then, we use a pretrained (and then adapted) text encoder to embed the set of $|\mathcal{Y}|$ potential labels into a continuous vector space $\mathbb{R}^C$ as learning objective and denote the language embeddings as $T \in \mathbb{R}^{|\mathcal{Y}| \times C}$. Next, given an input image

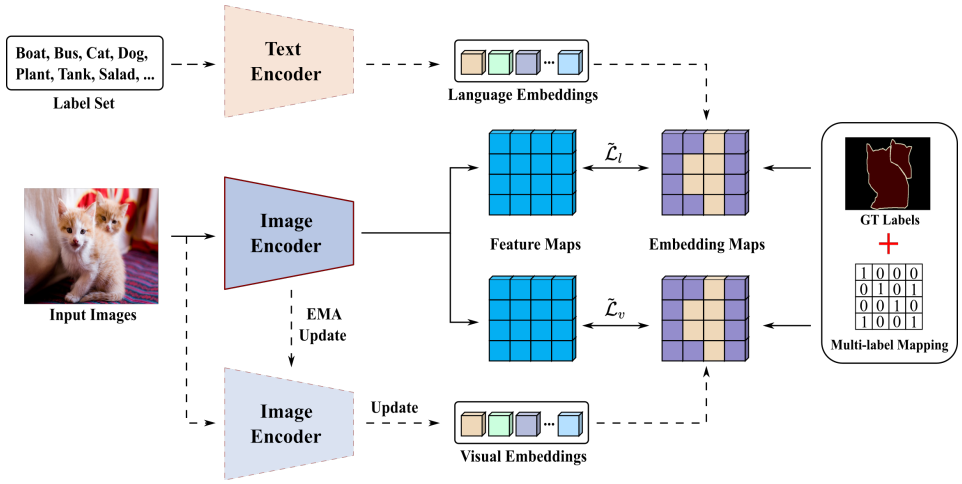

Figure 1: MS3 pipeline. Given the union segmentation dataset $\mathcal{D}$ with a combined label space $\mathcal{Y}$, we utilize a pretrained text encoder and a momentum updated image encoder to provide language and visual embeddings for every classes. Feature maps of the input images can be learned through two kinds of optimized pixel-embedding pairing losses with multi-label mapping obtained from Eq. 6.

with size $H \times W$, we send it to an image encoder (with a projection head) F and obtain a dense embedding map $I \in \mathbb{R}^{\tilde{H} \times \tilde{W} \times C}$, where $\tilde{H} = \frac{H}{s}, \tilde{W} = \frac{W}{s}$ and s is a downsampling factor determined by the encoder. For pixel i in I with ground-truth label y_i, we pull its embedding I_i close to its corresponding language embedding, and push it away from all other language embeddings, namely:

$$\mathcal{L}_l^i = -\mathbb{1}\left[y_i = j\right] \log \left(\frac{\exp\left(I_i \cdot T_j / \tau\right)}{\sum_{k=1}^{|\mathcal{Y}|} \exp\left(I_i \cdot T_k / \tau\right)} \right), \tag{1}$$

where τ is a pre-defined temperature parameter and we set it to 0.07 following LSeg (Li et al., 2022). The loss $\mathcal{L}_l$ is averaged over all pixels in I. Note that the text encoder is discarded after obtaining the language embeddings T while only the visual part is updated during the multi-dataset pretraining.

In addition to the unchanged language embeddings, we also use the image encoder to generate momentum-updated visual embeddings V for all the categories, which serve as another pixel-level feature learning objective. Similarly, we also pull the pixel embedding I_i close to its corresponding visual embedding and push it away from other visual embeddings:

$$\mathcal{L}_v^i = -\mathbb{1}\left[y_i = j\right] \log \left(\frac{\exp\left(I_i \cdot V_j / \tau\right)}{\sum_{k=1}^{|\mathcal{Y}|} \exp\left(I_i \cdot V_k / \tau\right)} \right). \tag{2}$$

The final loss $\mathcal{L}$ is a combination of $\mathcal{L}_l$ and $\mathcal{L}_v$:

$$\mathcal{L} = \mathcal{L}_l + \alpha \mathcal{L}_v, \tag{3}$$

where α controls the loss weight of the vision supervision and we simply set it to 1. While we experimentally confirm that the above straightforward multi-task loss can bring complementary improvements, this loss still has limitations in multi-dataset scenarios. As mentioned, we also conduct multi-label extension to $\mathcal{L}_l$ and $\mathcal{L}_v$ through a novel cross-modal information exchange module, which will be introduced in detail in Sect. 3.3.

3.2 Multimodal Supervision for Semantic Segmentation

Language supervision with improved embeddings. The text encoder of CLIP is powerful for associating text and image, and previous works (Li et al., 2022; Yin et al., 2022) usually choose it for

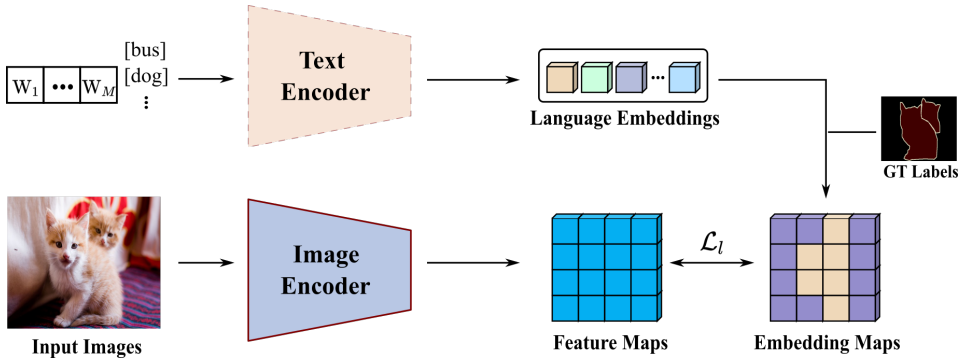

Figure 2: The pipeline for improving language embedding. We adapt the CLIP pretrained model to the segmentation datasets with learnable textual contexts as input and use the pixel-embedding pairing loss for feature learning. Then the adapted model can produce higher quality embeddings.

extracting text embeddings. However, these works mostly focus on zero shot learning, and we notice that directly using these embeddings is not suitable for our multi-dataset pretraining framework, for it introduces confused embeddings between classes. As shown in Tab. 1, since the original CLIP uses human-designed templates like "a photo of a [CLS]." as text prompts, the language embeddings mainly summarize this format information and lead to very high similarities between embeddings. Also, closely-connected class pairs like *'cat'* vs. *'dog'* and *'horse'* vs. *'cow'* exhibit highest similarity, and using the CLIP class embedding directly would make these classes confused.

To address this issue and better meet the mutli-dataset pretraining setting, we present a simple adaptive approach to alleviate the confusion and thus improve language embedding quality. As shown in Fig. 2, we adapt the CLIP pretrained model on the union segmentation dataset $\mathcal{D}$ with the following changes. First, inspired by CoOp (Zhou et al., 2021), we replace the human-designed templates with learnable textual contexts and use them as the input of the text encoder. The input of the text encoder then becomes $[\mathrm{W}]_1[\mathrm{W}]_2 \dots [\mathrm{W}]_M[\mathrm{CLS}]$, where $[\mathrm{W}]_m (m \in \{1, \dots, M\})$ represents M context tokens. Then, we change the image-embedding pairing objective to pixel-embedding pairing using Eq. 1. Finally, we use the adapted model to generate higher quality language embeddings for MS3. At this point, as shown in the last row of Tab. 1, the high similarity issue between different classes is alleviated, while the correlation between categories can still be reflected.

Vision supervision with momentum-updated embeddings. Considering that some categories which are indistinguishable by the language embeddings may be more visually discriminative, we additionally introduce visual embeddings as complements. Specifically, we send the input image to $\hat{F}$, which is a momentum updated version of vision encoder F, and obtain dense embedding $\hat{I}$. These embeddings are stored in a memory bank by categories for querying. Visual embedding for class j, V_j, is then calculated by the weighted average of all the pixels with label j in the memory bank:

$$V_j = \frac{\sum_{i=1}^{N_m} \mathbb{1}\left[y_i = j\right] W_{ij} M_i}{\sum_{i=1}^{N_m} \mathbb{1}\left[y_i = j\right]}, \tag{4}$$

where M_i and y_i are the pixel embedding and label for pixel i stored in the memory bank. N_m is the number of pixels in the memory bank and the weight W_{ij} is simply represented by the cosine similarity between M_i and the old visual embedding V_j^{old}. In this way, the visual embedding V is dynamically updated with the memory bank, and we conduct pixel-embedding pairing as in Eq. 4.

3.3 Cross-modal Multi-label Extension

Although the language and visual embeddings are helpful for modeling the relationship between categories, the one-by-one pixel-embedding correspondence of $\mathcal{L}_l$ and $\mathcal{L}_v$ still causes performance bottlenecks under the multi-dataset situation, since class embeddings with similar meanings in other datasets will be incorrectly pushed away. From the above consideration, we further provide

a multi-label extension for $\mathcal{L}_l$ and $\mathcal{L}_v$ by exchanging the similarity between visual features and language embeddings through a novel cross-modal information exchange module. Kim et al. (2022) also designs a multi-label extension strategy, but it is restricted to a single modal and can not well handle domain differences between datasets. While the advantages of our module come from the cross-modal alignment, which serves as an agent for reducing the domain gaps, thus introducing consistent cross-modal information for label mapping generation can make the generated mappings more comprehensive and reliable. Specifically, for class $y \in \mathcal{Y}_i$, we first calculate the mean cosine similarity of all the pixel embeddings belonging to this class with respect to the text embedding of class $y' \in \mathcal{Y}$ and denote it as $\mathcal{S}_y^{y'}$:

$$\mathcal{S}_y^{y'} = \frac{\sum_{k=1}^{N_p} \mathbb{1}\left[y_k = y\right] F_k \cdot T_{y'}}{\sum_{k=1}^{N_p} \mathbb{1}\left[y_k = y\right]}, \tag{5}$$

where F_k and y_k represent the embedding and the label of pixel k and N_p is the total pixel number. Then, the new multi-class label $\tilde{Y}_y \in \{0,1\}^{\mathcal{Y}}$ for class y is generated by:

$$\tilde{Y}_y^{y'} = \mathbb{1}[\mathcal{S}_y^{y'} \geq \mathcal{S}_y^{y} \text{ and } y' \in \mathcal{Y} \backslash \mathcal{Y}_i] \tag{6}$$

Note that the relation between y and y' can be unidirectional, *i.e.*, $\tilde{Y}_y^{y'} \neq \tilde{Y}_{y'}^{y}$, to deal with inclusion relationships between categories, *e.g.*, class y can be a subset of class y'. Finally, for pixel i with label y_i in the dense embedding map I, we not only pull it close to its corresponding language embedding and visual embedding, but also pull it close to the embeddings in $\tilde{Y}_{y_i}$, and the Eq. 1, Eq. 2 and Eq. 3 can be changed to:

$$\tilde{\mathcal{L}}_l^i = -\mathbb{1}\left[y_i = j\right] \log \left(\frac{\exp\left(I_i \cdot T_j / \tau\right) + \sum_{k=1}^{|\mathcal{Y}|} \mathbb{1}\left[\tilde{Y}_j^k\right] \exp\left(I_i \cdot T_k / \tau\right)}{\sum_{k=1}^{|\mathcal{Y}|} \exp\left(I_i \cdot T_k / \tau\right)} \right), \tag{7}$$

$$\tilde{\mathcal{L}}_v^i = -\mathbb{1}\left[y_i = j\right] \log \left(\frac{\exp\left(I_i \cdot V_j / \tau\right) + \sum_{k=1}^{|\mathcal{Y}|} \mathbb{1}\left[\tilde{Y}_j^k\right] \exp\left(I_i \cdot V_k / \tau\right)}{\sum_{k=1}^{|\mathcal{Y}|} \exp\left(I_i \cdot V_k / \tau\right)} \right), \tag{8}$$

$$\tilde{\mathcal{L}} = \tilde{\mathcal{L}}_l + \alpha \tilde{\mathcal{L}}_v. \tag{9}$$

4 EXPERIMENT

4.1 EXPERIMENTAL SETUPS

Datasets. Our experiments are conducted on five widely used segmentation benchmarks, VOC 2012 (Everingham et al., 2015), ADE20K (Zhou et al., 2019), COCO-Stuff (Caesar et al., 2018), Cityscapes (Cordts et al., 2016) and Mapillary (Neuhold et al., 2017). Details for these datasets are included in the appendix.

Implementation details. The framework is based on DeepLab-v3+ (Chen et al., 2018) with a standard ResNet-50 as backbone (He et al., 2016). In practice, we add two 3-layer projection heads after the ASPP layer of DeepLab-v3+, resulting in two 512-d dense embedding maps respectively for pairing language and visual embeddings. A SGD optimizer with momentum 0.9 and weight decay $4e-5$ is used for 100 epochs pretraining. The batch size and initial learning rate are set to 128 and 0.8, respectively, over 8 NVIDIA Tesla V100 GPUs. The learning rate is decayed to 0 by cosine scheduler (Loshchilov & Hutter., 2016). The input size is set to 224×224 for efficiency. For data augmentations, we choose random crop, color distortion, and Gaussian blur. For the language embedding improvement module, following Rao et al. (2022), we set M to 13.

During the fine-tuning, we follow the basic configuration of MMSegmentation[1] except using a standard ResNet-50 backbone and removing the auxiliary head for *conveniently comparing with*

[1] https://github.com/open-mmlab/mmsegmentation

Table 2: Standard fine-tuning results on five typical segmentation benchmarks. ⋆ means the methods are implemented and fit to our pretraining setting by ourselves.

Method	Pretrained Dataset				mIoU				
	ImageNet	VOC	ADE20K	COCO	VOC	ADE20K	COCO	Cityscapes	Mapillary
Scratch					44.78	28.67	25.02	54.27	22.23
MoCo-v2	✓				71.59	38.29	33.64	77.52	38.16
DenseCL	✓				72.68	38.12	33.77	77.63	37.71
PixPro	✓				75.37	39.34	34.87	78.24	39.16
Supervised	✓				75.63	39.36	35.25	77.60	38.21
LSeg⋆		✓	✓	✓	76.97	41.19	37.59	79.02	40.96
MDP⋆		✓	✓	✓	76.87	41.36	37.32	78.65	40.81
MS3		✓	✓	✓	**78.60**	**42.93**	**38.63**	**79.93**	**42.65**

other baselines and purely reflecting the influence of the backbone itself, while MS3 can still bring performance gains over stronger models, which is further discussed in the ablation study. The fine-tuning simply follow the common setting of each dataset, *i.e.,* for VOC, we fine-tune the pretrained model for $40k$ iterations using 513×513 input size, while for ADE20K and COCO-Stuff, the iterations are set to $80k$ with 512×512 input size. For Cityscapes and Mapillary, the iterations are set to $40k$ with 512×1024 and 768×768 input size, respectively.

Evaluation. We combine the training split of VOC, ADE20K and COCO-Stuff, and obtain around 150K training images with more than 300 classes for pretraining. The performance is evaluated on the three datasets to validate how multi-dataset pretraining boosts the performance. We also transfer the pretrained model to Cityscapes and Mapillary, where the model does not see during the pretraining stage to validate its generalization ability. Following the standard, we use mean Intersection-over-Union (mIoU) for performance evaluation.

4.2 Main Results

Standard fine-tuning. This section reports the standard fine-tuning results on five representative benchmarks. For a better understanding of the advantages of the proposed framework, we compare our results with the supervised ImageNet pretraining baseline (Supervised) and some typical self-supervised pretraining baselines, MoCo-v2 (Chen et al., 2020b), DenseCL (Wang et al., 2021c) and PixPro (Xie et al., 2020). We also fit LSeg (Li et al., 2022) and MDP (Shi et al., 2021) to our setting and list their results. From the results shown in Tab. 2, we find that:

- MS3 outperforms ImageNet pretraining baselines for a considerable margin. Specifically, our method achieves 78.60% mIoU on VOC, which is 7.01%, 5.92%, 3.23% and 2.97% better than MoCo-v2, DenseCL, PixPro and supervised ImageNet pretraining, respectively. For ADE20K and COCO-Stuff, the performance gains toward supervised ImageNet pretraining are 3.57% ($42.93\% vs. 39.36\%$) and 3.38% ($38.63\% vs. 35.25\%$), respectively.
- The performance advantages are held on two *unseen* domains, *i.e.*, Cityscapes and Mapillary. MS3 achieves 79.93% and 42.65% mIoU on these two benchmarks, which are 2.33% and 4.44% better than supervised ImageNet pretraining, respectively. The benefits also hold towards other baselines, indicating that our framework enjoys better transferability.

Frozen-backbone fine-tuning. We further try to freeze the pretrained image encoder and only adjust the classification head during the finetuning stage. As shown in Tab. 3, when freezing the backbone, our method can achieve 73.63%, 37.39% mIoU on VOC and ADE20K and obtain 36.24% mIoU on *unseen* Mapillary. The results far exceed freezing supervised ImageNet pretraining backbone, which demonstrates the effects of modeling inter-category relations and performing pixel-level discrimination during pretraining. We also notice that the performance of our framework can still maintain a high level under the freezing setting, which can meet the requirements of some application scenarios that require rapid adaptation.

Table 3: Results over VOC, ADE20K and Mapillary with frozen image encoder. † means the image encoder is frozen.

Method	mIoU		
	VOC	ADE20K	Mapillary
MoCo-v2	71.59	38.29	38.16
Supervised	75.63	39.36	38.21
MS3	**78.60**	**42.93**	**42.65**
MoCo-v2†	63.51	22.48	29.93
Supervised†	61.97	21.43	28.79
MS3†	**73.63**	**37.39**	**36.24**

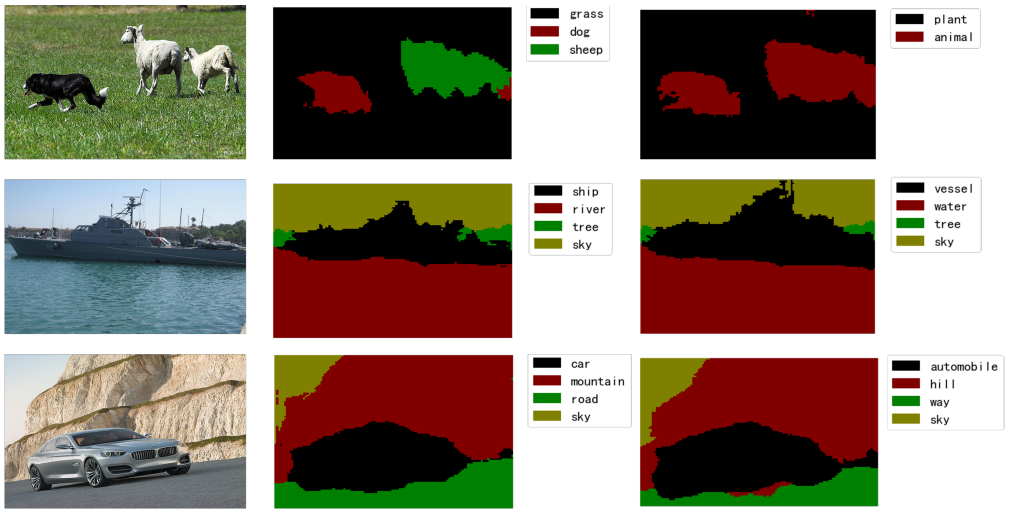

Figure 3: Zero-shot visualization. Our pretrained model can conduct some flexible segmentation with semantic similar labels or hierarchical labels.

Zero-shot prediction. Our pre-trained model can also be used for flexible zero-shot predictions, benefiting from the text embeddings. As shown in Fig. 3, given different category inputs, our pretrained model can obtain basically accurate segmentation results by matching pixel embeddings with the text embeddings of different categories. In addition to the synonymous semantic replacement of known categories, our model can make correct predictions for some categories that are not defined in pre-training (*e.g.*, *vessel* and *way*), and can also handle some hierarchical annotations.

4.3 Ablation Study

In this section, we conduct extensive ablation studies to verify the effectiveness of the proposed modules. *Unless specified, all models are pretrained over VOC and ADE20K for 100 epochs and evaluated on VOC for efficiency.*

Effects of multi-modal supervision. Tab. 4 verifies the effects of multi-modal supervision. In addition to the pixel-embedding pairing loss, we also list the results using cross-entropy loss (abbreviated as CE) under the multi-dataset scenario for comparison, where "single-head" means using one prediction head for all datasets while "multi-head" means using dataset-specific classifiers. We confirm that directly using CE loss performs poorly, especially under the single-head setting, due to class conflicts. Using the language embeddings obtained from the original CLIP pretrained model as supervision can bring some benefits, *i.e.*, 1.18% better than single-head CE and 0.84% multi-head CE, while it has no advantage towards our generated visual embedding ($71.93\% vs. 71.98\%$). The language embeddings and visual embeddings have complementary effects, and using them together achieves the best performance (72.60%).

Table 4: Effects of multi-modal supervision.

Method	mIoU
CE (single-head)	70.75
CE (multi-head)	71.09
Visual	71.98
Language	71.93
Visual + Language	**72.60**

Table 5: Benefits of improving language embedding quality. $\mathbb{M}_I$ represents the module for language embedding improvement.

Backbone	$\mathbb{M}_I$	dim	mIoU
VIT-B/32	✗	512	71.93
VIT-B/32	✓	512	72.65
ResNet50	✓	1024	**73.09**

Table 6: Effects of multi-label extension. $\mathbb{M}_C$ represents the cross-modal information exchange module. Note that we use the improved language embedding in this ablation.

Method	$\mathbb{M}_C$	mIoU
Visual	✗	71.98
Language	✗	72.65
Visual	✓	72.95
Language	✓	73.27
Visual + Language	✓	**73.68**

Table 7: Results when using better backbones and auxiliary loss.

Method	Backbone	mIoU	
		VOC	ADE20K
Supervised	Res50-v1c	76.81	42.72
Supervised	Res101-v1c	78.62	44.60
Supervised	DeiT-B	80.48	45.36
MS3	Res50-v1c	**78.92**	**44.05**
MS3	Res101-v1c	**79.81**	**45.41**
MS3	DeiT-B	**81.86**	**46.62**

Effects of improving language embeddings. Tab. 5 inspects the results when using the improved language embeddings, and the results are positive. The embeddings got through the adapted VIT-B backbone can bring 0.72% performance gain, from 71.93% to 72.65%. Higher performance can be achieved when using language embedding obtained from the adapted ResNet50 backbone, *i.e.*, 73.09% mIoU on VOC. We analyze that this benefit is due to the higher feature dimension of the generated embeddings, while considering the learning cost, we keep using 512-d embeddings in MS3.

Effects of cross-modal multi-label extension. Tab. 6 studies the effect of our proposed cross-modal information exchange module, which provides a multi-label extension to the loss function. The results are also promising, and we achieve 0.97% and 0.62% performance gains after providing multi-label maps for visual supervision and language supervision, respectively. The performance can be further boosted to 73.68% when compatible with all MS3 components including multimodal supervision, which demonstrates the superiority of our framework.

Results when equipped with better models We also evaluate MS3 with larger backbones and auxiliary loss. As shown in Tab. 7, when equipped with a Res50-v1c backbone and an auxiliary head, MS3 achieves 78.92% mIoU on VOC and 44.05% mIoU on ADE20K, which is respectively 2.11% and 1.33% better than supervised ImageNet pretraining, and the result is even comparable with Res101-v1c based supervised ImageNet pretraining results. When equipped with the Res101-v1c backbone, MS3 can further boost the performance to 79.81% and 45.41%, respectively. For transformer-based DeiT (Touvron et al., 2021) backbone[2] with UPerhead (Xiao et al., 2018) as decoder, our MS3 can still bring benefits, *i.e.*, 1.38% gains on VOC and 1.26% gains on ADE20K. The above results prove that our MS3 can bring consistent benefits when using larger capacity models.

5 Conclusion

This paper proposed a multi-dataset pretraining framework for semantic segmentation with multi-modal supervision. On one hand, we adapt CLIP pretrained text encoder to provide high-quality language embedding for each class. On the other, we generate momentum-updated visual embeddings as complements. These two kinds of embeddings are used together to guide feature learning in a pixel-embedding pairing way. We also additionally design a cross-modal information exchange module for multi-label loss extension for further improvement. Experiment results show that our framework consistently outperforms ImageNet pretrained models on several widely used benchmarks.

[2]For DeiT backbone, we use AdamW optimizer with 3e-4 learning rate for 100 epochs pretraining.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 833–851, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

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

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards open vocabulary object detection without human-provided bounding boxes. *arXiv preprint arXiv:2111.09452*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–998, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. *arXiv preprint arXiv:2202.14030*, 2022.

John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2876–2885, 2020.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiaodan Liang, Eric Xing, and Hongfei Zhou. Dynamic-structured semantic propagation network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 752–761, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1209–1218, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV*, 2017.

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. In *International Conference on Machine Learning (ICML)*, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)*, 115(3):211–252, 2015.

Bowen Shi, Xiaopeng Zhang, Haohang Xu, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. Multi-dataset pretraining: A unified model for semantic segmentation. *arXiv preprint arXiv:2202.02002*, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 10347–10357, July 2021.

Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation. *arXiv preprint arXiv:2103.11351*, 2021a.

Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021b.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021c.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020.

Wei Yin, Yifan Liu, Chunhua Shen, Anton van den Hengel, and Baichuan Sun. The devil is in the labels: Semantic segmentation from sentences. *arXiv preprint arXiv:2202.02002*, 2022.

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, et al. Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985*, 2020.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. *arXiv preprint arXiv:2112.09106*, 2021.

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 on Computer Vision (IJCV)*, 127:302–321, 2019.

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

Table 8: Examples of multi-label mapping. Note that the relation is unidirectional, *i.e.*, pixels in the former class of one class pairs can be mapped to the embedding of the latter class, while it does not hold conversely.

Class Pairs		Class Pairs	
Reliable relations:			
stairs	stairway	hill	mountain
land	earth	ship	boat
bulletin board	signboard	bar	counter
toaster	oven	hovel	horse
dirt track	path	lake	water
Typical mistakes:			
tree	vegetable	fountain	waterfall
grass	gravel	sand	sea

A Dataset Details

As mentioned, our experiments are based on five segmentation benchmarks, namely:

- **VOC 2012** (Everingham et al., 2015) contains 10,582 training (including the annotations provided by (Hariharan et al., 2011)), 1,449 validation, and 456 test images with pixel level annotations for 20 foreground object classes and one background class.
- **ADE20K** (Zhou et al., 2019) is a scene parsing dataset. It contains around 25K images spanning 150 semantic categories, of which 20K for training, 2K for validation, and another 3K for testing.
- **COCO-Stuff** (Caesar et al., 2018) is a large scale dataset, which includes 118K training images and 5k validation images from COCO 2017 (Lin et al., 2014). It provides rich annotations for 80 object classes and 91 stuff classes.
- **Cityscapes** (Cordts et al., 2016) is consist of 5,000 finely annotated urban scene images, with 2,975/500/1,524 for train/val/test, respectively. Its performance is reported on 19 challenging categories.
- **Mapillary** (Neuhold et al., 2017) is a street-level imagery dataset with pixel-accurate and instance-specific human annotation. It contains 25K high-resolution images annotated into 66 object categories.

B Examples for Multi-label Mapping

This section inspects the generated multi-label mapping for our cross-modal interactive loss. As shown in Tab. 8, our multi-label mapping can correctly reflect some similar semantics, *e.g.*, *'stairs' vs. 'stairway'* and *'hill' vs. 'mountain'*. Furthermore, due to the unidirectional characteristic of the generated mapping, our method can also handle the annotation granularity problem well, *e.g.*, class *'dirty track'* is a subset of class *'path'*, so pixels with label *'dirty track'* should be pulled close to the embedding of *'path'*, while it does not hold conversely. The above relationships are correctly represented in our generated mappings, but inevitably, there are also some errors in our mapping,

Table 9: Comparing with single dataset pretraining on VOC.

Pretraining Dataset	Size	mIoU
VOC	10K	69.59
ADE20K	20K	69.90
COCO	118K	77.10
VOC and ADE20K	30K	**73.68**
VOC, ADE20K and COCO	148K	**78.60**

including incorrectly mapping some semantically similar but different class pairs (*e.g.*, *'fountain' vs. 'waterfall'*) and some class pairs that usually appear in the same scene at the same time (*e.g.*, *'sand' vs. 'sea'*). Among them, we believe that the reason for the latter mistake is mainly because our framework uses a small input resolution, and expanding the input resolution can alleviate it to a certain extent.

C Compared with Single Dataset Pretraining

We conduct single dataset pretraining using our MS3 framework to verify the necessity of multi-dataset pretraining. As shown in Tab. 9, MS3 achieves 69.59% and 69.90% mIoU on VOC, respectively, when separately using VOC and ADE20K for pretraining, and combining them for multi-dataset pretraining can boost the performance to 73.68%. We also find that the dataset size counts for MS3, and merely using the larger COCO-Stuff can achieve 77.10% mIoU, while MS3 can still benefit from introducing other small-size datasets, and combining all three datasets for pretraining further brings a 1.50% performance gain (78.60%).

D Adding Unlabeled Data for Pretraining

In addition to labeled datasets, our framework can utilize multi-modal embeddings to provide reliable pseudo labels for unlabelled data for further pretraining. Specifically, for pixel i in the dense embedding map I_u of an unlabeled input, we calculate the cosine similarity of its pixel embedding I_u^i to all the language embeddings T and all the visual embeddings V. If the most similar language embedding and visual embedding meet the same category, we regard such pseudo labeling as reliable and assign pixel i with the corresponding class.

To verify the effectiveness of the above scheme, we use labeled VOC and ADE20K and add *unlabeled* COCO-Stuff data for jointly pretraining. The results are shown in Tab. 10. Using additional unlabeled data can bring 1.09% performance gain, from 73.68% to 74.77%, which proves that our framework has the ability to utilize large-scale unlabeled data for further improvement.

Table 10: Studies on adding unlabeled data. $\mathcal{D}$ contains VOC and ADE20K. $\mathcal{U}$ is *unlabeled* COCO-Stuff.

$\mathcal{D}$	$\mathcal{U}$	mIoU
✓	✗	73.68
✓	✓	**74.77**