cases	doc_1		doc_2		decision	id
	authors	• Chaofan Ma • Yuhuan Yang • Yanfeng Wang • Ya Zhang • Weidi Xie	**authors**	• Ma, Chaofan • Wang, Yanfeng • Xie, Weidi • Yang, Yuhuan • Zhang, Ya	DUPLICATES	170
	title	Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models	**title**	Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models		
	publication_date	2022-10-27 00:00:00	**publication_date**	2022-10-26 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.CORE		
	journal		**journal**			
	volume		**volume**			
	doi		**doi**	None		
	urls	• https://web.archive.org/web/20221029045449/https://arxiv.org/pdf/2210.15138v1.pdf	**urls**	• http://arxiv.org/abs/2210.15138		
	id	id4605788917229352841	**id**	id-2081344633025645888		
	abstract	When trained at a sufficient scale, self-supervised learning has exhibited a notable ability to solve a wide range of visual or language understanding tasks. In this paper, we investigate simple, yet effective approaches for adapting the pre-trained foundation models to the downstream task of interest, namely, open-vocabulary semantic segmentation. To this end, we make the following contributions: (i) we introduce Fusioner, with a lightweight, transformer-based fusion module, that pairs the frozen visual representation with language concept through a handful of image segmentation data. As a consequence, the model gains the capability of zero-shot transfer to segment novel categories; (ii) without loss of generality, we experiment on a broad range of self-supervised models that have been pre-trained with different schemes, e.g. visual-only models (MoCo v3, DINO), language-only models (BERT), visual-language model (CLIP), and show that, the proposed fusion approach is effective to any pair of visual and language models, even those pre-trained on a corpus of uni-modal data; (iii) we conduct thorough ablation studies to analyze the critical components in our proposed Fusioner, while evaluating on standard benchmarks, e.g. PASCAL-5i and COCO-20i , it surpasses existing state-of-the-art models by a large margin, despite only being trained on frozen visual and language features; (iv) to measure the model's robustness on learning visual-language correspondence, we further evaluate on synthetic dataset, named Mosaic-4, where images are constructed by mosaicking the samples from FSS-1000. Fusioner demonstrates superior performance over previous models.	**abstract**	When trained at a sufficient scale, self-supervised learning has exhibited a notable ability to solve a wide range of visual or language understanding tasks. In this paper, we investigate simple, yet effective approaches for adapting the pre-trained foundation models to the downstream task of interest, namely, open-vocabulary semantic segmentation. To this end, we make the following contributions: (i) we introduce Fusioner, with a lightweight, transformer-based fusion module, that pairs the frozen visual representation with language concept through a handful of image segmentation data. As a consequence, the model gains the capability of zero-shot transfer to segment novel categories; (ii) without loss of generality, we experiment on a broad range of self-supervised models that have been pre-trained with different schemes, e.g. visual-only models (MoCo v3, DINO), language-only models (BERT), visual-language model (CLIP), and show that, the proposed fusion approach is effective to any pair of visual and language models, even those pre-trained on a corpus of uni-modal data; (iii) we conduct thorough ablation studies to analyze the critical components in our proposed Fusioner, while evaluating on standard benchmarks, e.g. PASCAL-5i and COCO-20i , it surpasses existing state-of-the-art models by a large margin, despite only being trained on frozen visual and language features; (iv) to measure the model's robustness on learning visual-language correspondence, we further evaluate on synthetic dataset, named Mosaic-4, where images are constructed by mosaicking the samples from FSS-1000. Fusioner demonstrates superior performance over previous models.Comment: BMVC 2022 Ora		
	versions		**versions**			