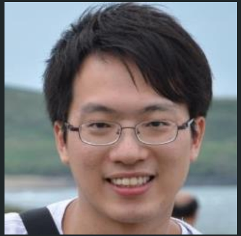
Open Vocabulary Semantic Segmentation with Patch Aligned Contrastive Learning

THU-AM-279


Jishnu Mukhoti [1,2]


Tsung-Yu Lin [1]


Omid Poursaeed [1]


Rui Wang [1]

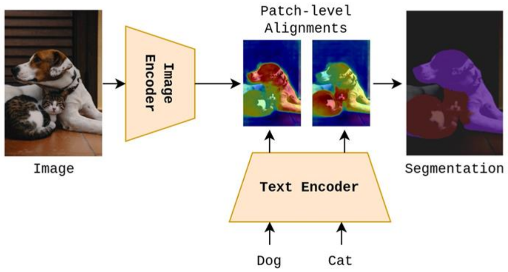

Ashish Shah [1]

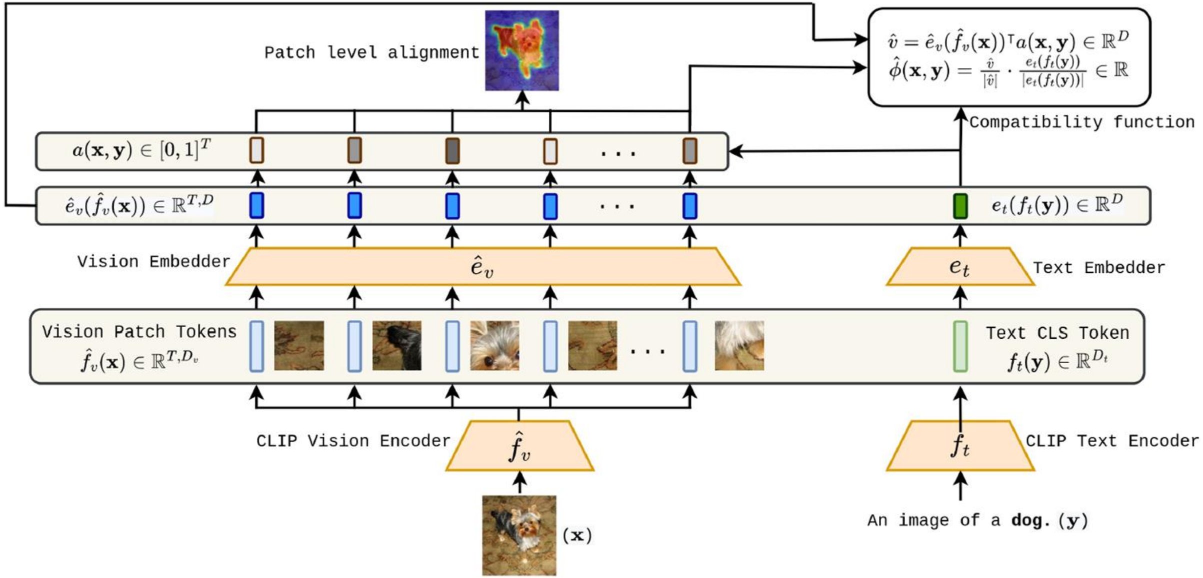

Philip H.S. Torr [2]


Ser-Nam Lim [1]

1 - Meta AI
2 - Torr Vision Group (TVG), University of Oxford

Patch Aligned Contrastive Learning (PACL)





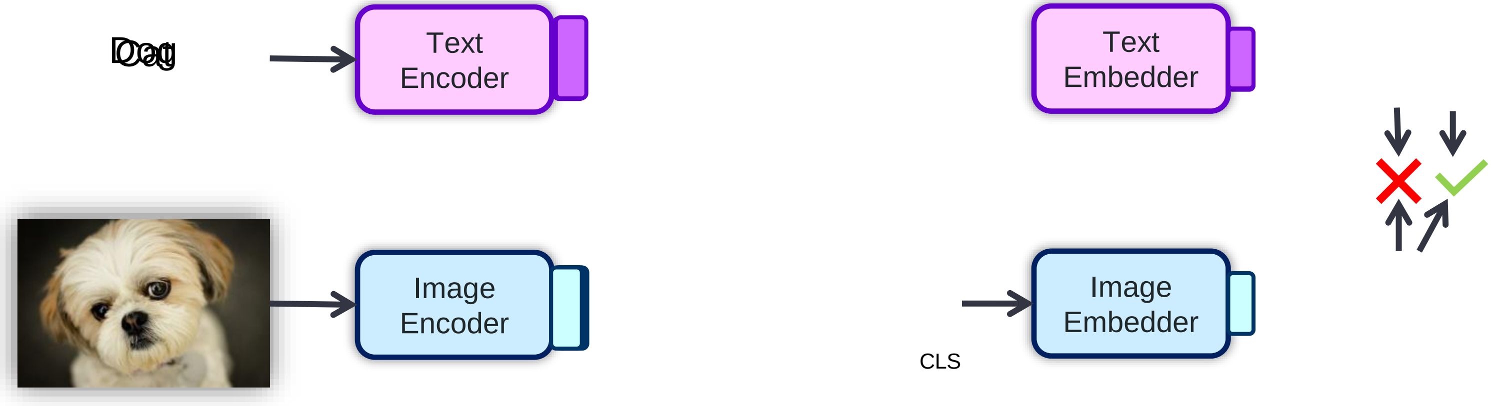
Semantic Segmentation

We label and predict on every pixel. This makes collecting annotations very expensive and limits the number of concepts that we can learn.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

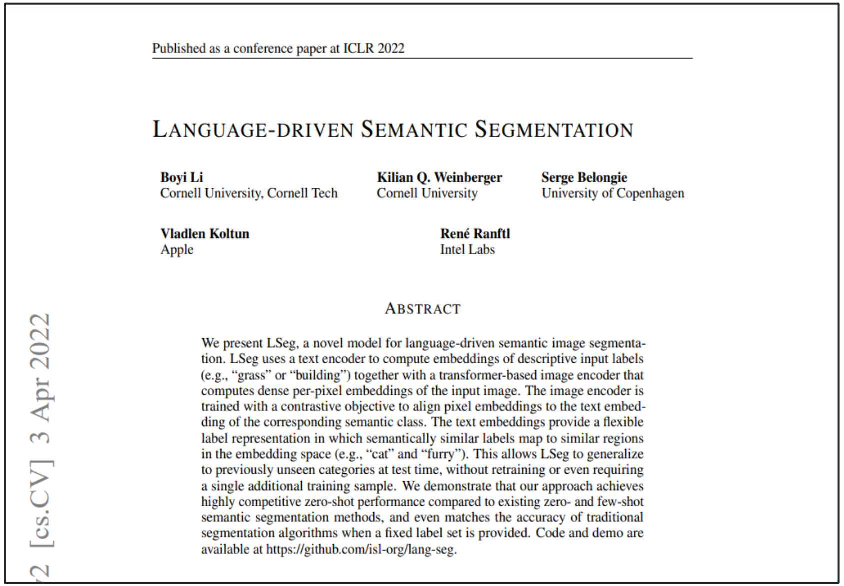
Open Vocabulary Prediction

Open vocabulary prediction in vision involves recognizing any arbitrary concept in an image. The concept can be described in natural language.



Like classification, can open-vocabulary semantic segmentation be performed using CLIP/CLIP like models?

Prior Work



Published as a conference paper at ICLR 2022

LANGUAGE-DRIVEN SEMANTIC SEGMENTATION

Boyi Li
Cornell University, Cornell Tech

Kilian Q. Weinberger
Cornell University

Serge Belongie
University of Copenhagen

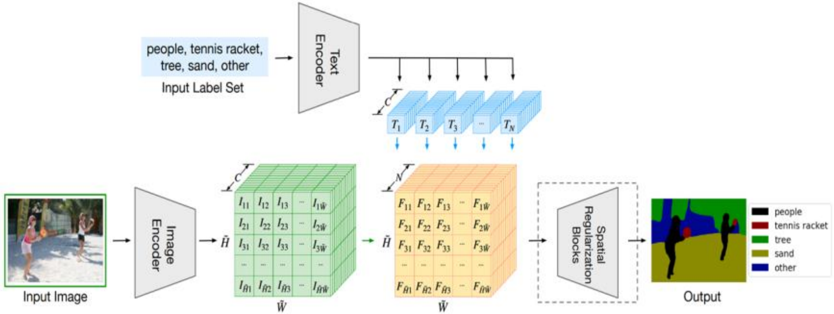
Vladlen Koltun
Apple

René Ranftl
Intel Labs

ABSTRACT

We present LSeg, a novel model for language-driven semantic image segmentation. LSeg uses a text encoder to compute embeddings of descriptive input labels (e.g., "grass" or "building") together with a transformer-based image encoder that computes dense per-pixel embeddings of the input image. The image encoder is trained with a contrastive objective to align pixel embeddings to the text embedding of the corresponding semantic class. The text embeddings provide a flexible label representation in which semantically similar labels map to similar regions in the embedding space (e.g., "cat" and "furry"). This allows LSeg to generalize to previously unseen categories at test time, without retraining or even requiring a single additional training sample. We demonstrate that our approach achieves highly competitive zero-shot performance compared to existing zero- and few-shot semantic segmentation methods, and even matches the accuracy of traditional segmentation algorithms when a fixed label set is provided. Code and demo are available at https://github.com/isl-org/lang-seg.

2 [cs.CV] 3 Apr 2022



Li, B., Weinberger, K.Q., Belongie, S., Koltun, V. and Ranftl, R., 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

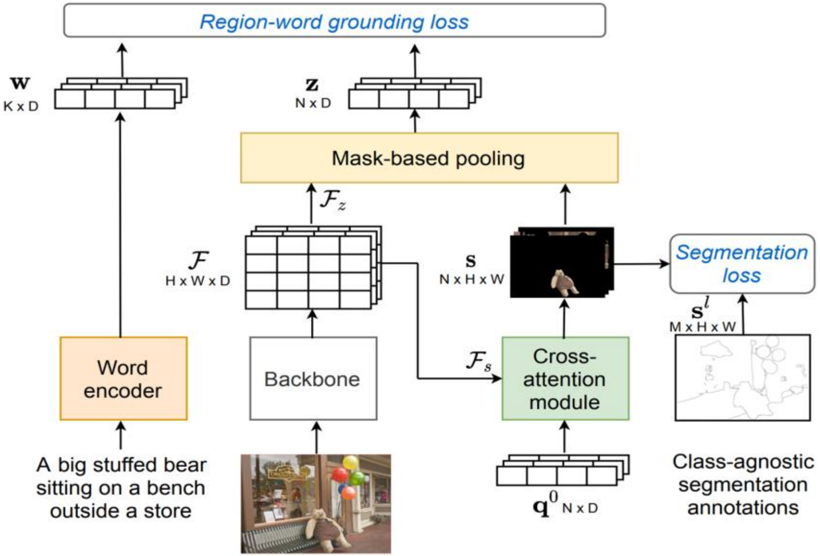
Prior Work

Scaling Open-Vocabulary Image Segmentation with Image-Level Labels

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin*

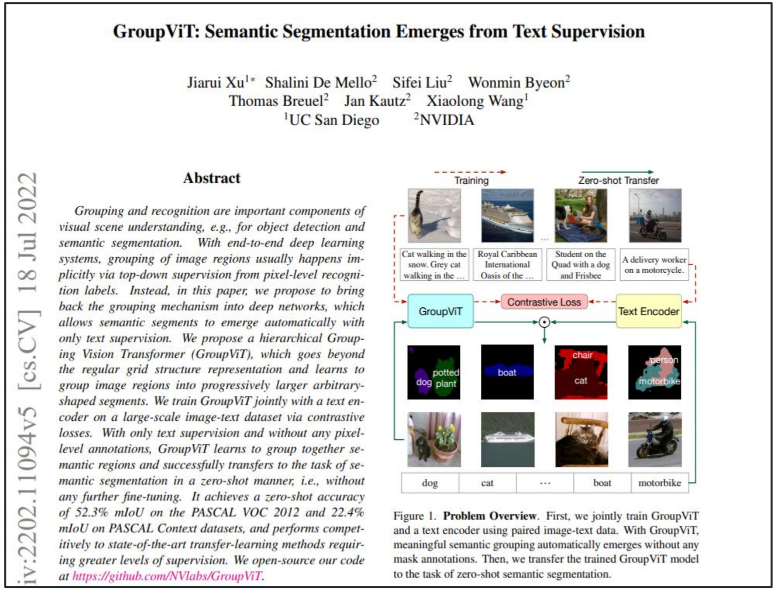
Google Research
{golnazg, xiuyegu, yincui}@google.com tsungyil@nvidia.com

Abstract We design an open-vocabulary image segmentation model to organize an image into meaningful regions indicated by arbitrary texts. Recent works (CLIP and ALIGN), despite attaining impressive open-vocabulary classification accuracy with image-level caption labels, are unable to segment visual concepts with pixels. We argue that these models miss an important step of visual grouping, which organizes pixels into groups before learning visual-semantic alignments. We propose OpenSeg to address the above issue while still making use of scalable image-level supervision of captions. First, it learns to propose segmentation masks for possible organizations. Then it learns visual-semantic alignments by aligning each word in a caption to one or a few predicted masks. We find the mask representations are the key to support learning image segmentation from captions, making it possible to scale up the dataset and vocabulary sizes. OpenSeg significantly outperforms the recent open-vocabulary method of LSeg by +19.9 mIoU on PASCAL dataset, thanks to its scalability.



Ghiasi, G., Gu, X., Cui, Y. and Lin, T.Y., 2021. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*.

Prior Work



GroupViT: Semantic Segmentation Emerges from Text Supervision

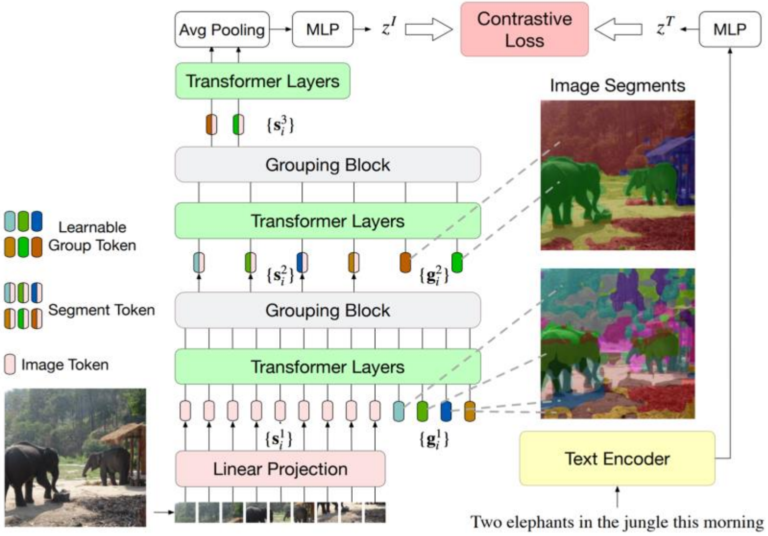
Jiarui Xu[1*] Shalini De Mello[2] Sifei Liu[2] Wonmin Byeon[2]
Thomas Breuel[2] Jan Kautz[2] Xiaolong Wang[1]
[1]UC San Diego [2]NVIDIA

arXiv:2202.11094v5 [cs.CV] 18 Jul 2022

Abstract

Grouping and recognition are important components of visual scene understanding, e.g., for object detection and semantic segmentation. With end-to-end deep learning systems, grouping of image regions usually happens implicitly via top-down supervision from pixel-level recognition labels. Instead, in this paper, we propose to bring back the grouping mechanism into deep networks, which allows semantic segments to emerge automatically with only text supervision. We propose a hierarchical Grouping Vision Transformer (GroupViT), which goes beyond the regular grid structure representation and learns to group image regions into progressively larger arbitrary-shaped segments. We train GroupViT jointly with a text encoder on a large-scale image-text dataset via contrastive losses. With only text supervision and without any pixel-level annotations, GroupViT learns to group together semantic regions and successfully transfers to the task of semantic segmentation in a zero-shot manner, i.e., without any further fine-tuning. It achieves a zero-shot accuracy of 52.3% mIoU on the PASCAL VOC 2012 and 22.4% mIoU on PASCAL Context datasets, and performs competitively to state-of-the-art transfer-learning methods requiring greater levels of supervision. We open-source our code at https://github.com/NVlabs/GroupViT.

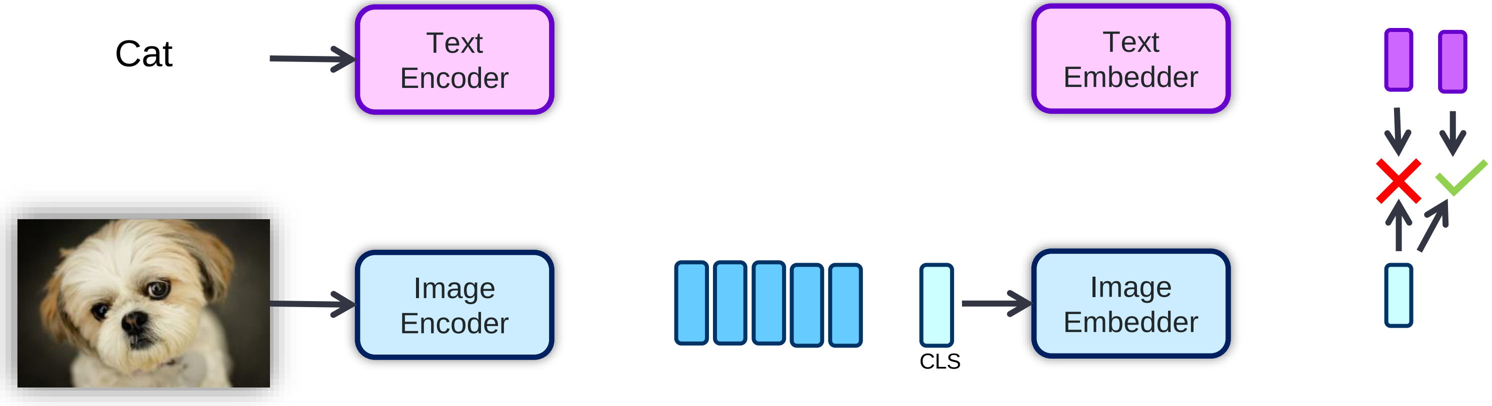
Figure 1. **Problem Overview.** First, we jointly train GroupViT and a text encoder using paired image-text data. With GroupViT, meaningful semantic grouping automatically emerges without any mask annotations. Then, we transfer the trained GroupViT model to the task of zero-shot semantic segmentation.



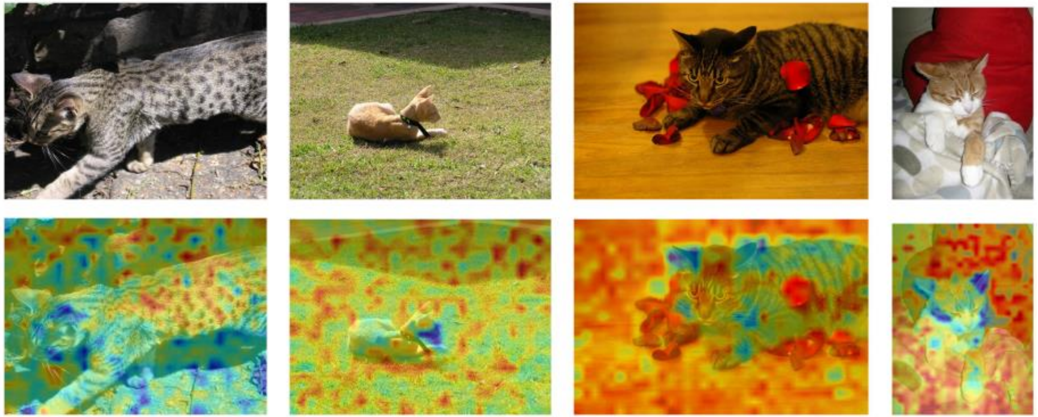
(a) GroupViT Architecture and Training Pipeline

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J. and Wang, X., 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18134-18144).

What if we use the patch vision tokens?



Patch-token level alignment in CLIP



Alignment of different images of cats with the word "cat". We see similar results even when providing prompt engineered versions of the word "cat".

CLIP Vision Encoder	Patch Classification Accuracy *Pre-Alignment*	*Post-Alignment*
ViT-B-16	52.49	96.51
ViT-L/14	27.91	95.33

Using the alignment value to classify patches leads to extremely poor classification accuracy.

In short, we see no patch/token level alignment between the image and text encoders in a pre-trained CLIP.

Patch-token level alignment in CLIP

Is there any semantically useful information in the vision patch tokens then?

Semantic Coherence

Semantic Coherence: The property which leads to semantically similar regions in the image having similar patch/token level representations.

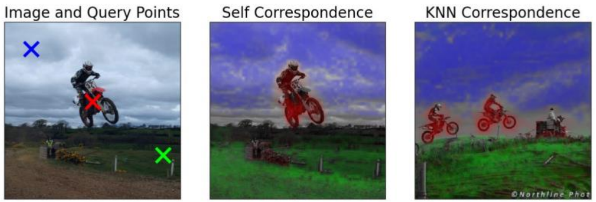


Figure 2: Feature correspondences from DINO. Correspondences between the source image (left) and the target images (middle and right) are plotted over the target images in the respective color of the source point (crosses in the left image). Feature correspondences can highlight key aspects of shared semantics within a single image (middle) and across similar images such as KNNs (right)

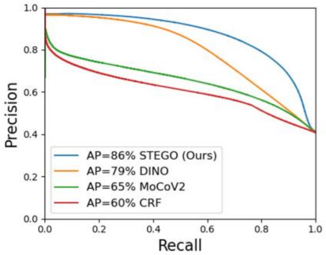


Figure 3: Precision recall curves show that feature self-correspondences strongly predict true label co-occurrence. DINO outperforms MoCoV2 and a CRF kernel, which shows its power as an unsupervised learning signal.

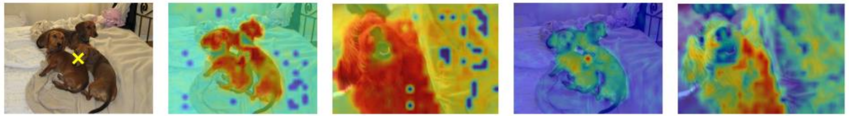
Semantic coherence has been shown to exist in pre-trained self-supervised vision transformers like DINO. Particularly, *feature correspondences can be used as a binary classifier to predict class-cooccurrence.*

In fact, this feature has been utilized before to train models for unsupervised semantic segmentation.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N. and Freeman, W.T., 2022. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*.

Semantic Coherence in CLIP

Semantic Coherence: The property which leads to semantically similar regions in the image having similar patch/token level representations.



(a) (b) (c) (d) (e)

Figure 4. **Qualitative results on semantic coherence between CLIP and DINO ViT-B/16. a)**: we show the original image of a dog class with the patch marker (yellow X near the centre). **b, c)**: we show CLIP vision encoder cosine similarity across all patches for the same and a different image of a dog. **d, e)**: we show the same for DINO. See more examples in Appendix B.1.

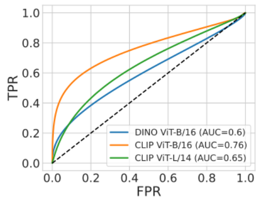
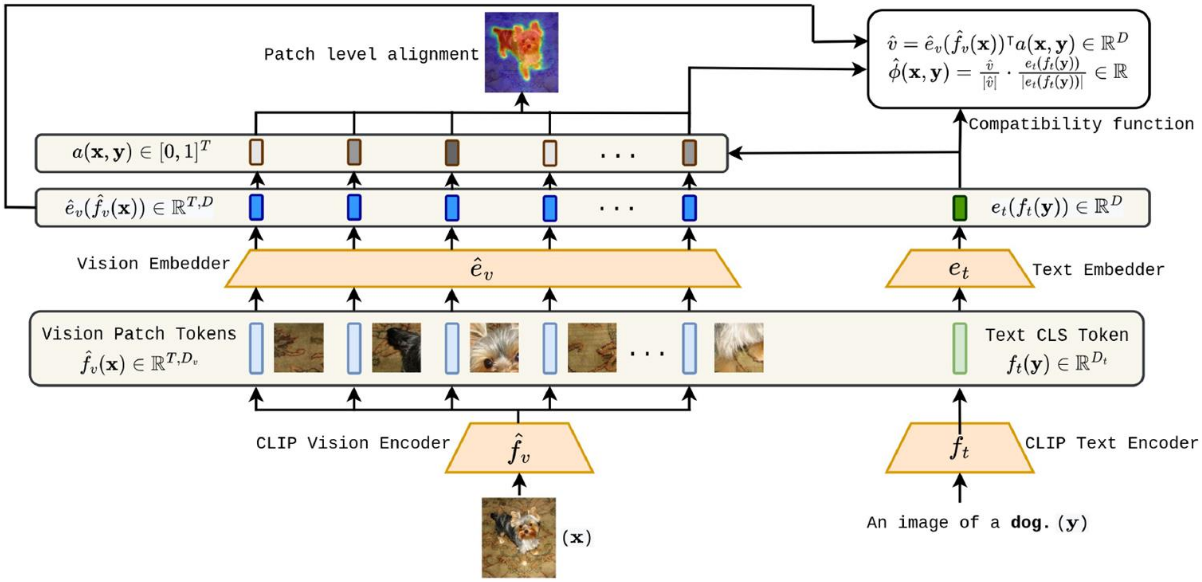


Figure 3. **ROC curve indicating semantic coherence of CLIP and DINO vision encoders.** CLIP encoders outperform DINO.

1. *Not only do we find semantic coherence to exist, we find the coherence to be stronger than DINO in terms of predicting class-cooccurrence.*
2. *Surprisingly, ViT-B/16 shows better coherence than ViT-L/14 in CLIP.*
3. *This indicates that we may be able to train an alignment between image and text patches.*
4. *This however has to be done in a **weakly supervised fashion**!*

Patch Aligned Contrastive Learning (PACL)



Alignment in CLIP vs CLIP + PACL

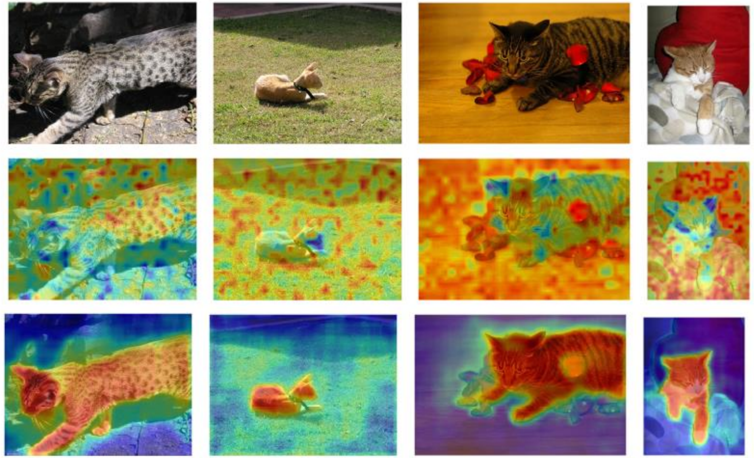
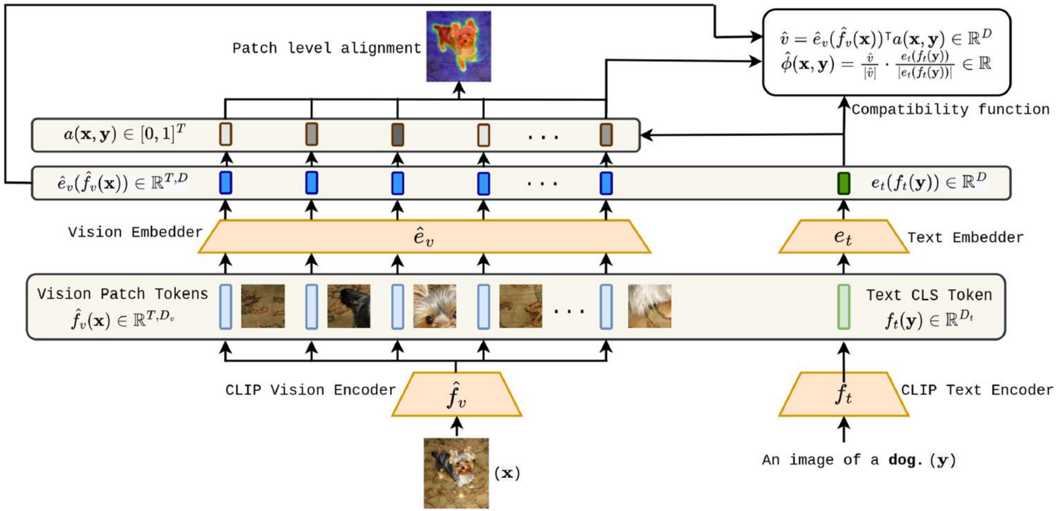


Figure 2. **Patch level alignment between the word "cat" and images of cats.** In the first row, we show the original images, in the second row, we show the patch level alignment in CLIP ViT-B/16 and in the third row, we show the alignment for our method.

CLIP Vision Encoder	Patch Classification Accuracy	
	Pre-Alignment	*Post-Alignment*
ViT-B-16	52.49	96.51
ViT-L/14	27.91	95.33

Zeroshot Segmentation and Classification



1. *Use the patch level alignment for segmentation.*
2. *Use the image level compatibility function for classification.*

Open Vocabulary Segmentation using PACL with CLIP Backbone

Method	Encoder	External Training Set	Constraints		mIoU			
			Annotation	*Mask*	*PV-20* [16]	*PC-59* [36]	*CS-171* [4]	*A-150* [63]
SPNet [55]	ResNet-101	✗	✓	✗	15.6	4.0	8.7	-
ZS3Net [3]	ResNet-101	✗	✓	✗	17.7	7.7	9.6	-
LSeg [28]	ViT-L/16	✗	✓	✗	52.3	-	-	-
OpenSeg [17]	EfficientNet-B7	COCO [9] + Loc. Narr. [40]	✗	✓	72.2	48.2	-	28.6
ViL-Seg [32]	ViT-B/16	GCC12M [6]	✗	✗	34.4	16.3	16.4	-
GroupViT [56]	ViT-S/16	GCC12M [6] + YFCC15M [41, 46]	✗	✗	52.3	22.4	24.3	-
CLIP [41]	ViT-B/16	WIT-400M [41]	✗	✗	8.4	2.3	2.6	1.3
CLIP + PACL (Ours)	ViT-B/16	GCC3M [44] + GCC12M [6] + YFCC15M [41, 46]	✗	✗	**72.3**	**50.1**	**38.8**	**31.4**

Table 2. **Results on zero-shot semantic segmentation** on Pascal VOC (PV-20), Pascal Context (PC-59) and COCO Stuff (CS-171) and ADE20K (A-150) datasets. We provide the encoder architecture, external training dataset (if any) as well as if those methods use segmentation annotations or class-agnostic segmentation masks. Our method (CLIP + PACL) consistently outperforms all previous approaches.

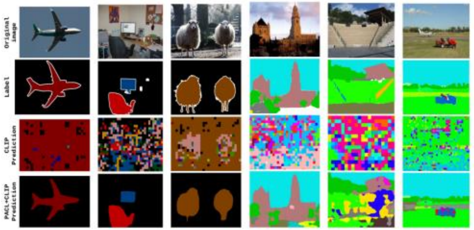


Figure 6. **Qualitative results on zero-shot semantic segmentation.** The first row denotes the original images, the second row shows the corresponding labels, the third row shows results obtained from a vanilla CLIP ViT-B/16, and the fourth row shows results of our method, PACL trained on a CLIP ViT-B/16 encoder. The first 3 images from the left are from Pascal VOC and the next 3 images are from ADE20K.

Stride Trick: *At inference time, reduce the stride of the convolutional layer generating patches in the transformer.*

This leads to a much larger number of patches and can be used to provide fine-grained segmentation predictions.

Ablations across Datasets and Encoders

Dataset	Vision Encoder	Text Encoder	mIoU PV-20
GCC12M	CLIP B/16	B/16	64.1
	CLIP L/14	L/14	62.7
	DINO B/16	B/16	55.4
GCC12M + YFCC15M	CLIP B/16	B/16	69.2
	CLIP L/14	L/14	68.4
	DINO B/16	B/16	62.6
GCC3M + GCC12M + YFCC15M	**CLIP B/16**	**B/16**	**72.3**
	CLIP L/14	L/14	71.7
	DINO B/16	B/16	64.8

Table 3. **Ablation on zero-shot segmentation across encoder architectures and datasets** on Pascal VOC (PV-20). In the Text Encoder column, B/16(L/14) indicates the pre-trained text encoder trained for CLIP ViT-B/16(L/14).

1. *The order of performance seems to follow the same trend as semantic coherence.*
2. *A patch level alignment can even be trained between DINO and a CLIP text encoder.*
3. *Thus PACL is independent of encoders and can be trained using different encoder combinations.*

Zero-shot Image Classification

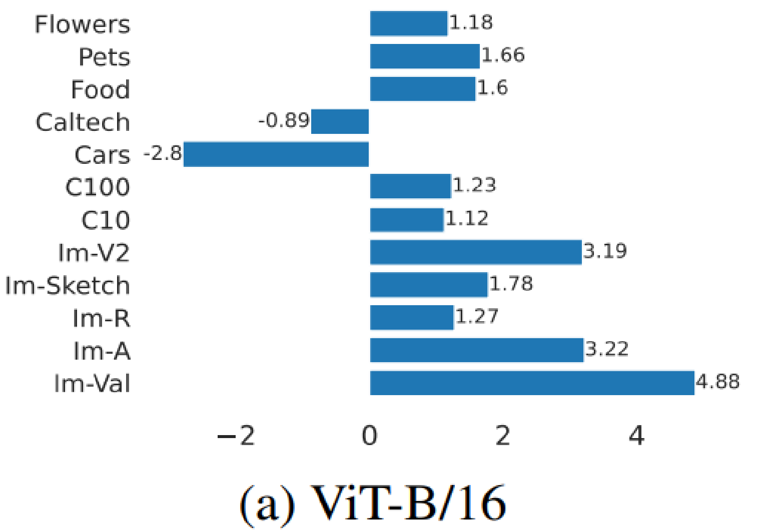


Figure 8. **Zero-shot image classification performance of PACL + CLIP vs vanilla CLIP on 12 datasets.** PACL + CLIP is competitive with or outperforms CLIP on most datasets.

References

[1] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

[2] Li, B., Weinberger, K.Q., Belongie, S., Koltun, V. and Ranftl, R., 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

[3] Ghiasi, G., Gu, X., Cui, Y. and Lin, T.Y., 2021. Open-vocabulary image segmentation. arXiv preprint arXiv:2112.12143.

[4] Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N. and Freeman, W.T., 2022. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*.

Thank you!