# CLIPSELF: VISION TRANSFORMER DISTILLS ITSELF FOR OPEN-VOCABULARY DENSE PREDICTION

**Size Wu**[1] **Wenwei Zhang**[1] **Lumin Xu**[2]
**Sheng Jin**[3,4] **Xiangtai Li**[1] **Wentao Liu**[4,5] **Chen Change Loy**[1]
[1] S-Lab, Nanyang Technological University [2] The Chinese University of Hong Kong
[3] The University of Hong Kong [4] SenseTime Research and Tetras.AI [5] Shanghai AI Laboratory
size001@e.ntu.edu.sg ccloy@ntu.edu.sg

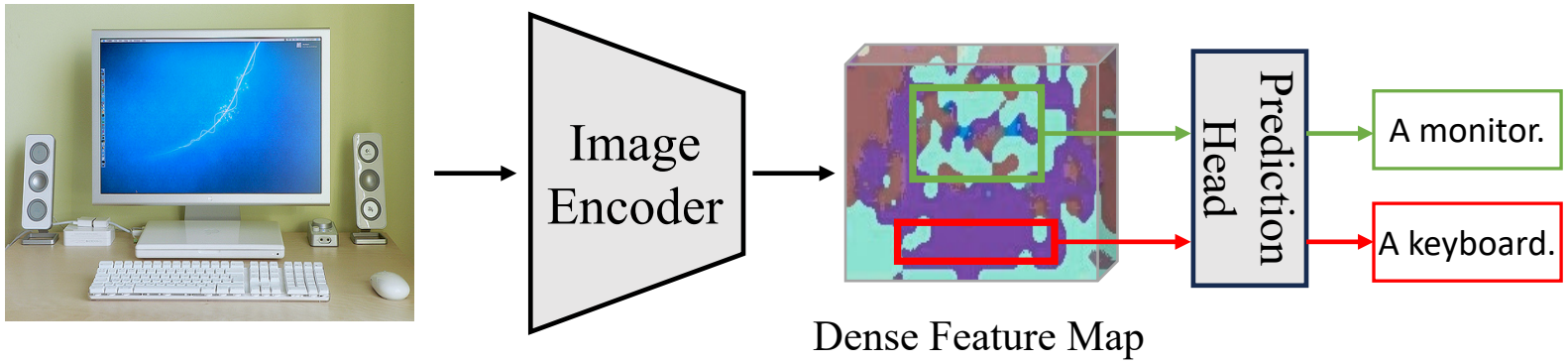
ICLR 2024

# Outline

- Open-Vocabulary Dense Prediction
- CLIP Image and Dense Representation
- CLIPSelf
- Experiments
- Visualization & Analysis

# Open-Vocabulary Dense Prediction

- Detect / Segment any object categories described by text

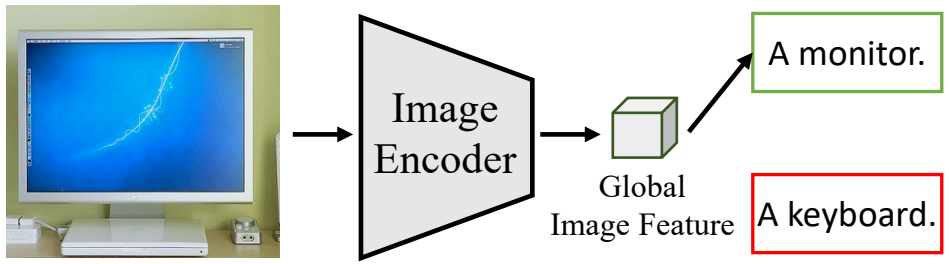


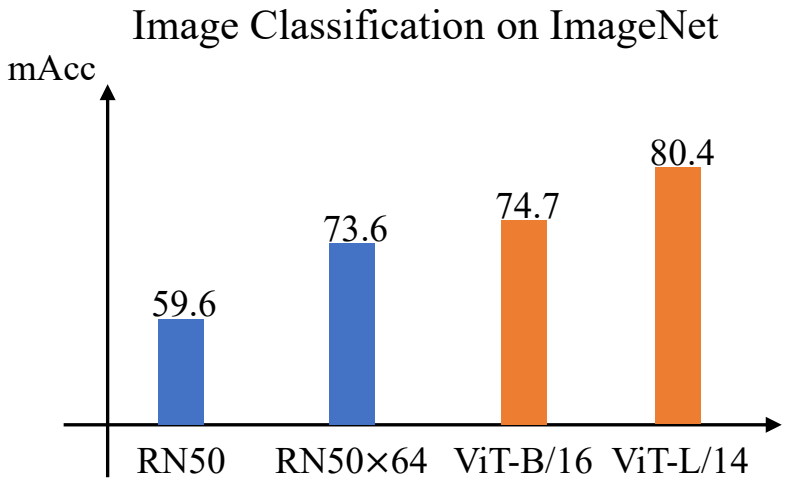
# Outline

- Open-Vocabulary Dense Prediction
- CLIP Image and Dense Representation
- CLIPSelf
- Experiments
- Visualization & Analysis

# CLIP Image and Dense Representation

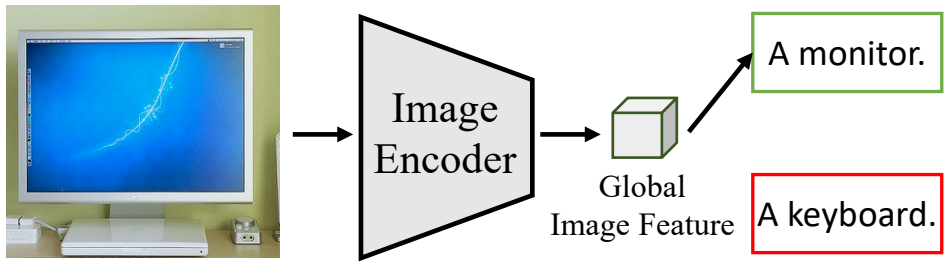
- Image Recognition by CLIP Image Representation
  - Pretrained on large-scale image-text pairs
  - Strong zero-shot image recognition ability

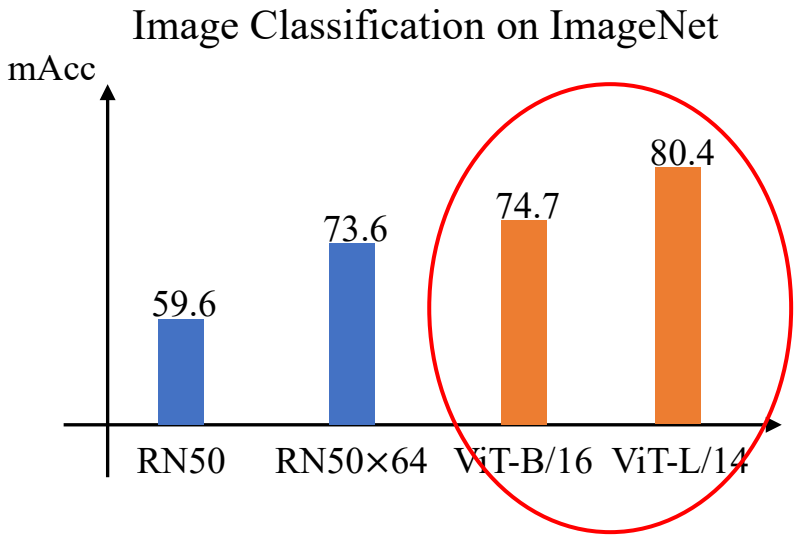


# CLIP Image and Dense Representation

- Image Recognition by CLIP Image Representation
  - Pretrained on large-scale image-text pairs
  - Strong zero-shot image recognition ability
  - The **ViT-based variants** exhibit superiority



# CLIP Image and Dense Representation

- Region Recognition
  a) Crop the regions and classify the **image crops** using CLIP image representation
  b) Extract region features from the **dense feature** map of the whole image

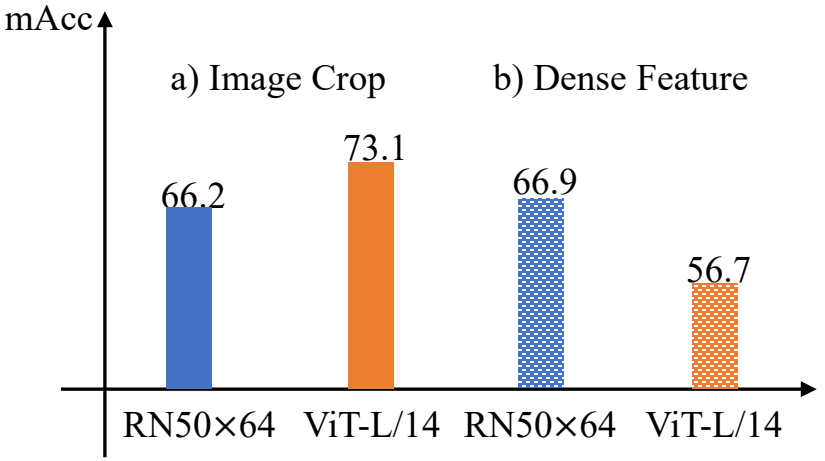

a) Image Crop

b) Dense Feature

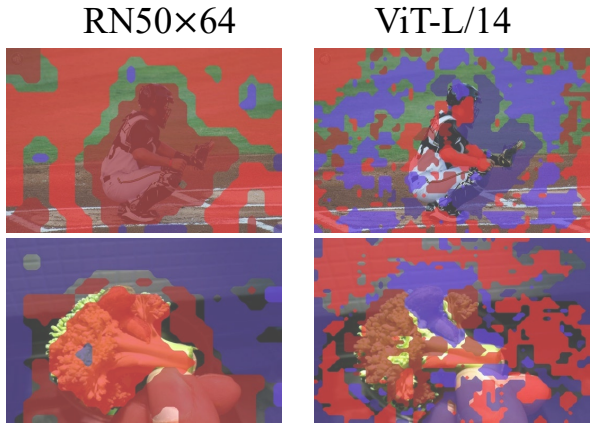
# CLIP Image and Dense Representation

- Region Recognition
  a) Crop the regions and classify the **image crops** using CLIP image representation
  b) Extract region features from the **dense feature** map of the whole image

Region Classification on COCO



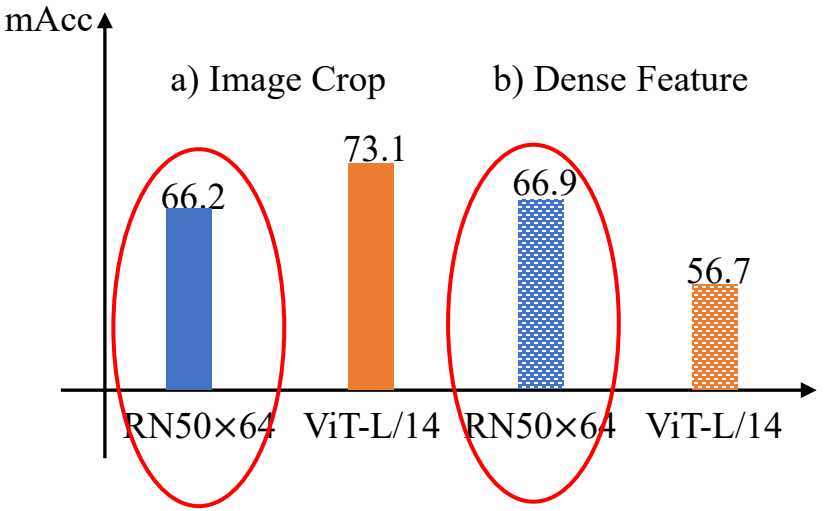

K-Means Cluster of Dense Feature



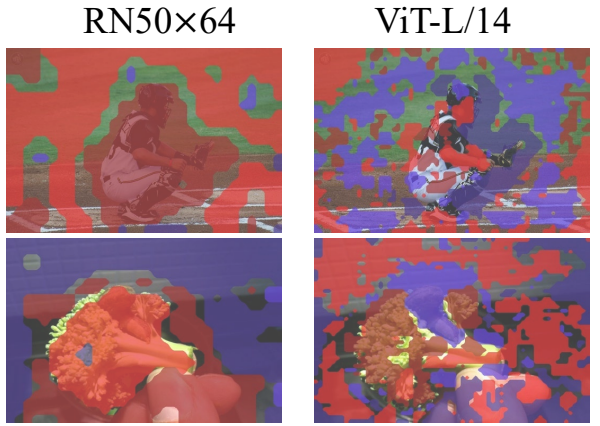
# CLIP Image and Dense Representation

- Region Recognition
  - The dense features of CNN-base models exhibit strong zero-shot ability for region recognition, even outperform image representations

Region Classification on COCO



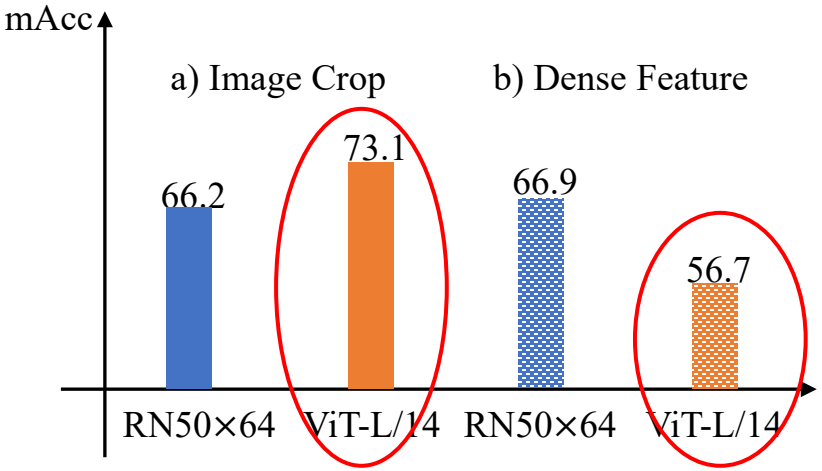

K-Means Cluster of Dense Feature



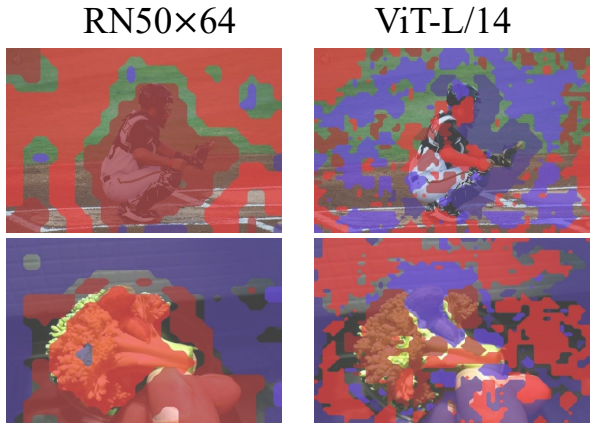
# CLIP Image and Dense Representation

- Region Recognition
  - The dense features of CNN-base models exhibit strong zero-shot ability for region recognition, even outperform image representations
  - The dense features of CLIP ViTs exhibit poor region recognition ability, while using image crops achieves satisfactory results

Region Classification on COCO



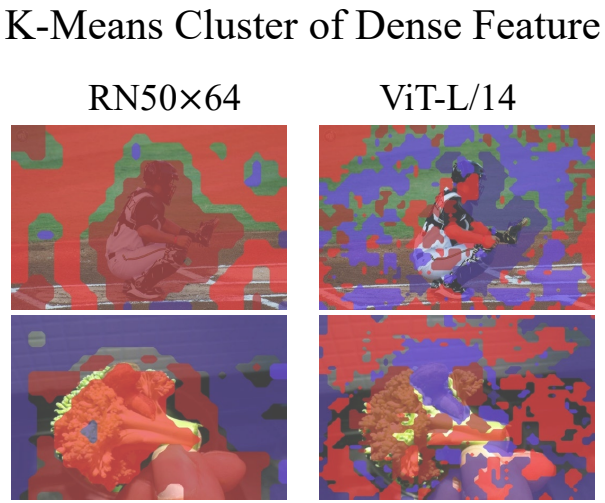

K-Means Cluster of Dense Feature



# CLIP Image and Dense Representation

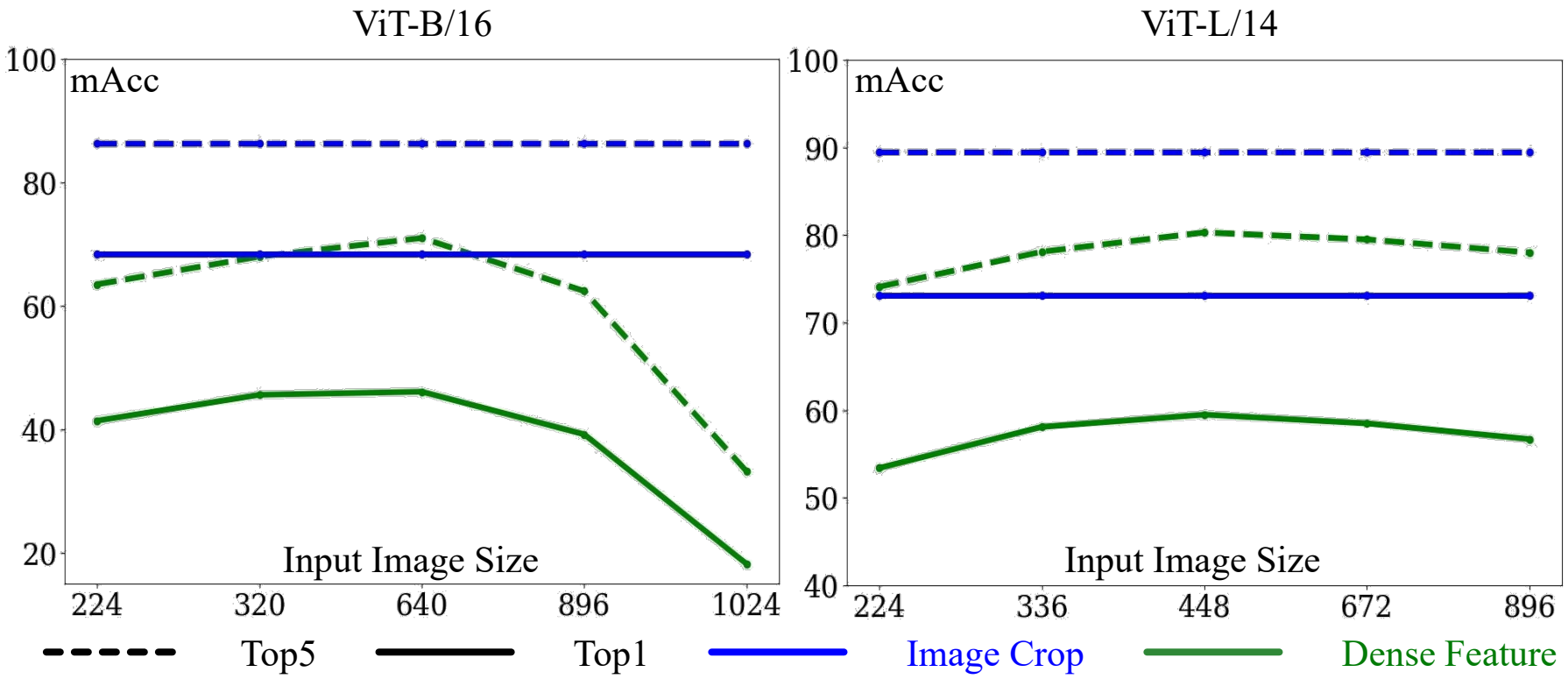
- Region Recognition
  - The K-Means results of feature map also illustrate that the CNN-based model better preserve the locality of image features



# CLIP Image and Dense Representation

- Region Recognition
  - A more comprehensive comparison between CLIP ViTs' Image Representation and Dense Representation

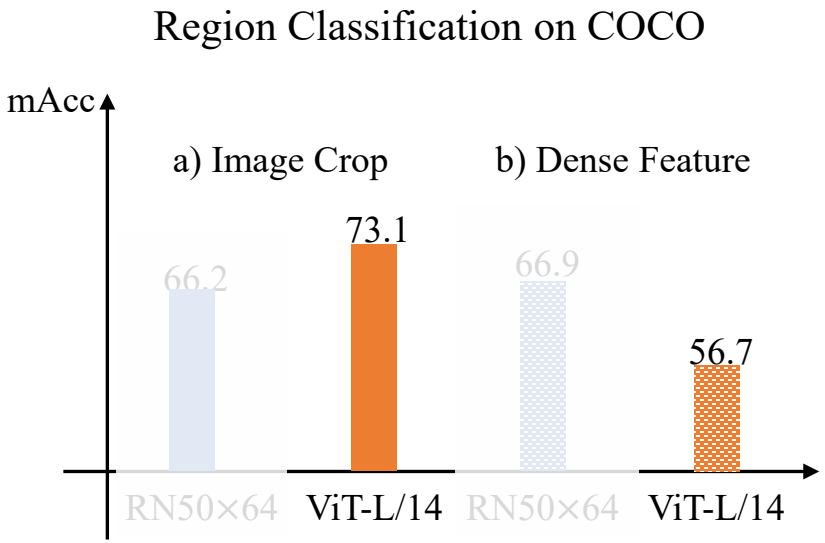


# Outline

- Open-Vocabulary Dense Prediction
- CLIP Image and Dense Representation
- CLIPSelf
- Experiments
- Visualization & Analysis

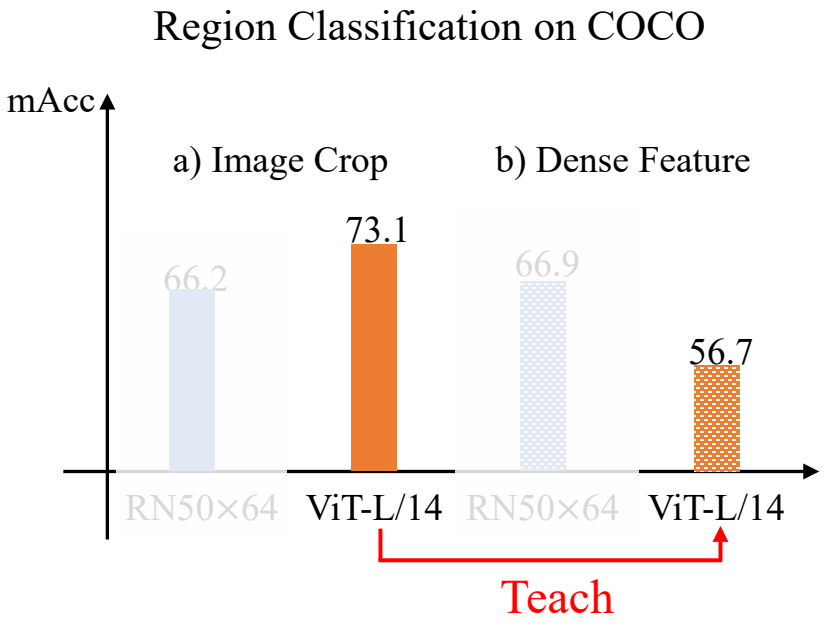
# CLIPSelf

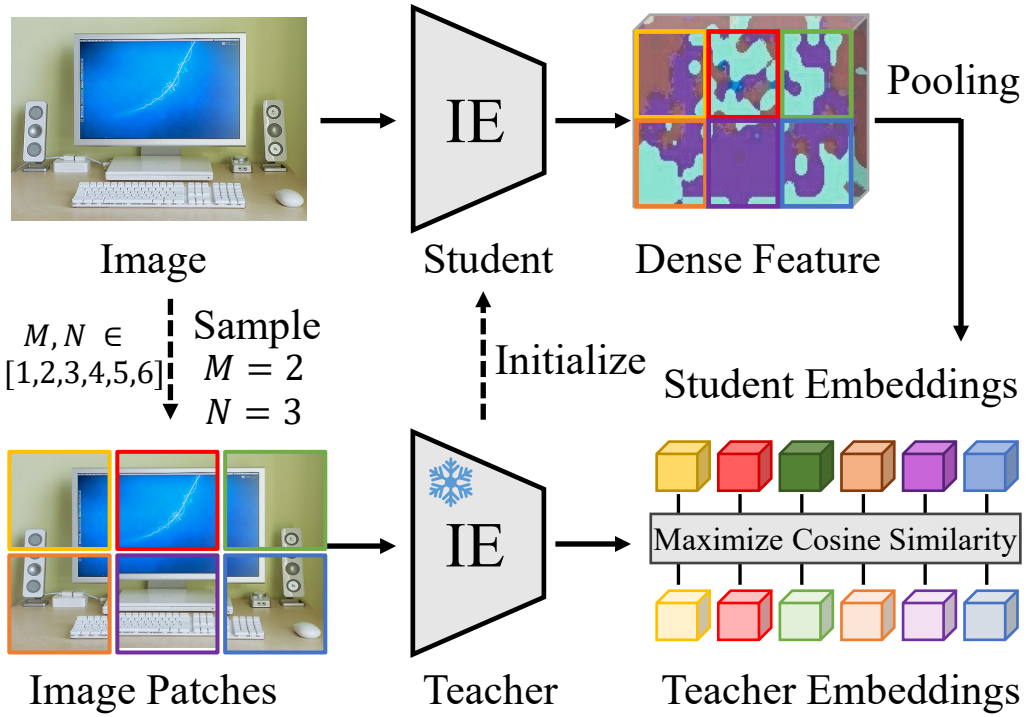
- Motivation: Enhancing the CLIP ViTs' dense representation
  1. The CLIP ViTs have exhibit powerful image representation, and they can effectively recognize the image crops of the object regions.
  2. How can we adapt the image-level recognition ability to CLIP ViTs' dense representation?



# CLIPSelf

- CLIPSelf: a self-distillation approach
  - The image representation of the cropped regions acts as the **teacher**
  - The region representation extracted from the dense feature map act as the **student**



# Outline

- Open-Vocabulary Dense Prediction
- CLIP Image and Dense Representation
- CLIPSelf
- Experiments
- Visualization & Analysis

# Experiments

- Quantitative results on region recognition

Table 2: Enhancement of dense representation. We report the Top1 and Top5 mean accuracy on classifying boxes and panoptic masks (thing and stuff).

| # | Model | Method | Region Proposals | Boxes | | Thing Masks | | Stuff Masks | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Top1 | Top5 | Top1 | Top5 | Top1 | Top5 |
| 1 | ViT-B/16 | - | - | 18.2 | 33.2 | 20.6 | 36.5 | 18.4 | 43.5 |
| 2 | ViT-B/16 | CLIPSelf | ✗ | 72.1 | 91.3 | 74.4 | 91.8 | **46.8** | **80.2** |
| 3 | ViT-B/16 | CLIPSelf | ✓ | **74.0** | **92.6** | **76.3** | **92.8** | 36.8 | 75.0 |

# Experiments

- Quantitative results on open-vocabulary dense prediction tasks

Table 3: Results on open-vocabulary object detection. 'L', 'B' and 'H' in ViT-based methods stand for base, large and huge model sizes. '/16' and '/14' stand for the downsample ratio of input images.

(a) OV-COCO benchmark

| Method | Backbone | $AP_{50}^{novel}$ |
|---|---|---|
| ViLD (Gu et al., 2021) | RN50 | 27.6 |
| Detic (Zhou et al., 2022b) | RN50 | 27.8 |
| F-VLM (Kuo et al., 2023) | RN50 | 28.0 |
| OV-DETR (Zang et al., 2022) | RN50 | 29.4 |
| BARON-KD (Wu et al., 2023a) | RN50 | 34.0 |
| CORA (Wu et al., 2023c) | RN50x4 | 41.7 |
| CORA+ (Wu et al., 2023c) | RN50x4 | 43.1 |
| PromptOVD (Song & Bang, 2023) | ViT-B/16 | 30.6 |
| RO-ViT (Kim et al., 2023b) | ViT-L/16 | 33.0 |
| CFM-ViT (Kim et al., 2023a) | ViT-L/16 | 34.1 |
| F-ViT | ViT-B/16 | 17.5 |
| F-ViT+CLIPSelf | ViT-B/16 | 37.6 |
| F-ViT | ViT-L/14 | 24.7 |
| F-ViT+CLIPSelf | ViT-L/14 | **44.3** |

(b) OV-LVIS benchmark

| Method | Backbone | $mAP_r$ |
|---|---|---|
| ViLD (Gu et al., 2021) | RN50 | 16.6 |
| OV-DETR (Zang et al., 2022) | RN50 | 17.4 |
| BARON-KD (Wu et al., 2023a) | RN50 | 22.6 |
| CORA+ (Wu et al., 2023c) | RN50x4 | 28.1 |
| F-VLM (Kuo et al., 2023) | RN50x64 | 32.8 |
| PromptOVD (Song & Bang, 2023) | ViT-B/16 | 23.1 |
| OW-ViT (Minderer et al., 2022) | ViT-L/14 | 25.6 |
| RO-ViT (Kim et al., 2023b) | ViT-L/16 | 32.4 |
| CFM-ViT (Kim et al., 2023a) | ViT-L/16 | 33.9 |
| RO-ViT (Kim et al., 2023b) | ViT-H/16 | 34.1 |
| F-ViT | ViT-B/16 | 11.5 |
| F-ViT+CLIPSelf | ViT-B/16 | 25.3 |
| F-ViT | ViT-L/14 | 24.2 |
| F-ViT+CLIPSelf | ViT-L/14 | **34.9** |

# Experiments

- Quantitative results on open-vocabulary dense prediction tasks

Table 4: Results on open-vocabulary semantic segmentation.

| Method | Model | ADE-150 | | ADE-847 | | PASCAL Context | |
|---|---|---|---|---|---|---|---|
| | | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc |
| SAN (Xu et al., 2023b) | ViT-B/16 | 27.5 | 45.6 | 10.1 | 21.1 | 53.8 | 73.0 |
| SAN (Xu et al., 2023b) | ViT-L/14 | 32.1 | 50.7 | **12.4** | 25.2 | 57.7 | 77.6 |
| Cat-Seg (Cho et al., 2023) | ViT-B/16 | 27.2 | 41.2 | 8.4 | 16.6 | 57.5 | 74.0 |
| Cat-Seg (Cho et al., 2023) | ViT-L/14 | 31.5 | 46.2 | 10.8 | 20.5 | 62.0 | 78.3 |
| Cat-Seg+CLIPSelf | ViT-B/16 | 29.0 | 46.0 | 9.3 | 20.1 | 58.0 | 75.3 |
| Cat-Seg+CLIPSelf | ViT-L/14 | **34.5** | **54.8** | **12.4** | **25.4** | **62.3** | **80.7** |

# Experiments

- Quantitative results on open-vocabulary dense prediction tasks

Table 5: Results on open-vocabulary panoptic segmentation. † means the results are obtained by running ODISE's officially released code and model.

| Method | Model | Score | | COCO Panoptic | | | ADE20K | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CLIP | Pred | PQ | mAP | mIoU | PQ | mAP | mIoU |
| ODISE (Xu et al., 2023a)† | ViT-L/14 | ✓ | ✗ | 27.6 | 26.2 | 23.7 | 15.3 | 9.8 | 17.3 |
| ODISE (Xu et al., 2023a)† | ViT-L/14 | ✓ | ✓ | 45.3 | 38.1 | **52.3** | 22.9 | 13.4 | 28.5 |
| ODISE+CLIPSelf | ViT-L/14 | ✓ | ✗ | 35.1 | 30.9 | 36.7 | 19.5 | 10.6 | 24.5 |
| ODISE+CLIPSelf | ViT-L/14 | ✓ | ✓ | **45.7** | **38.5** | **52.3** | **23.7** | **13.6** | **30.1** |

# Outline

- Open-Vocabulary Dense Prediction
- CLIP Image and Dense Representation
- CLIPSelf
- Experiments
- Visualization & Analysis

# Visualization & Analysis

- K-Means visualization of the enhanced dense representation

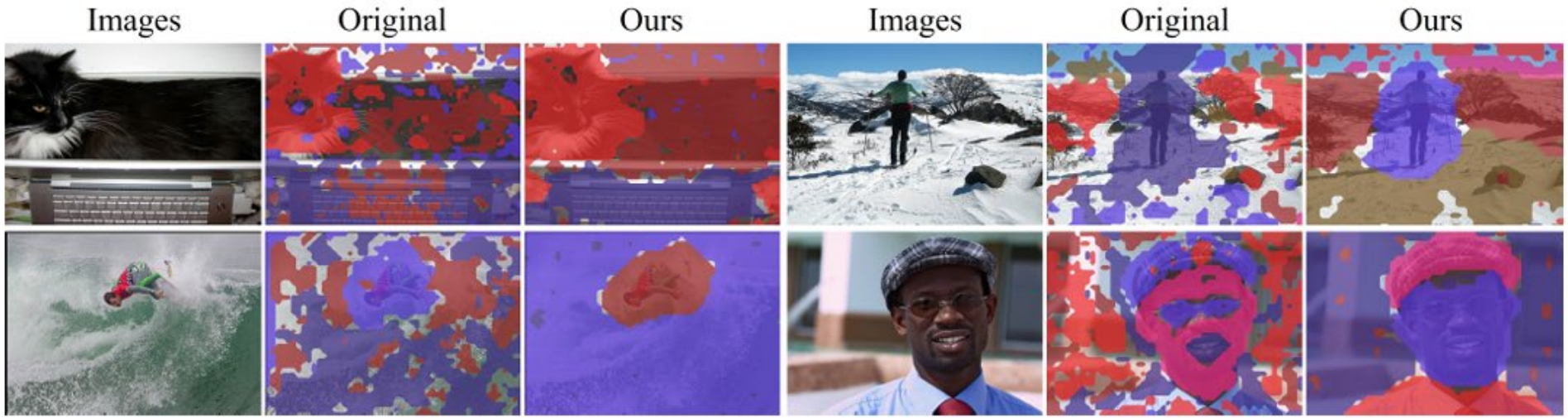



Figure 4: K-Means visualization of the dense feature maps of CLIP ViT. We show the raw images, the K-Means results of the original model, and those of our fine-tuned model by CLIPSelf.

# Visualization & Analysis

- Qualitative results on dense prediction tasks

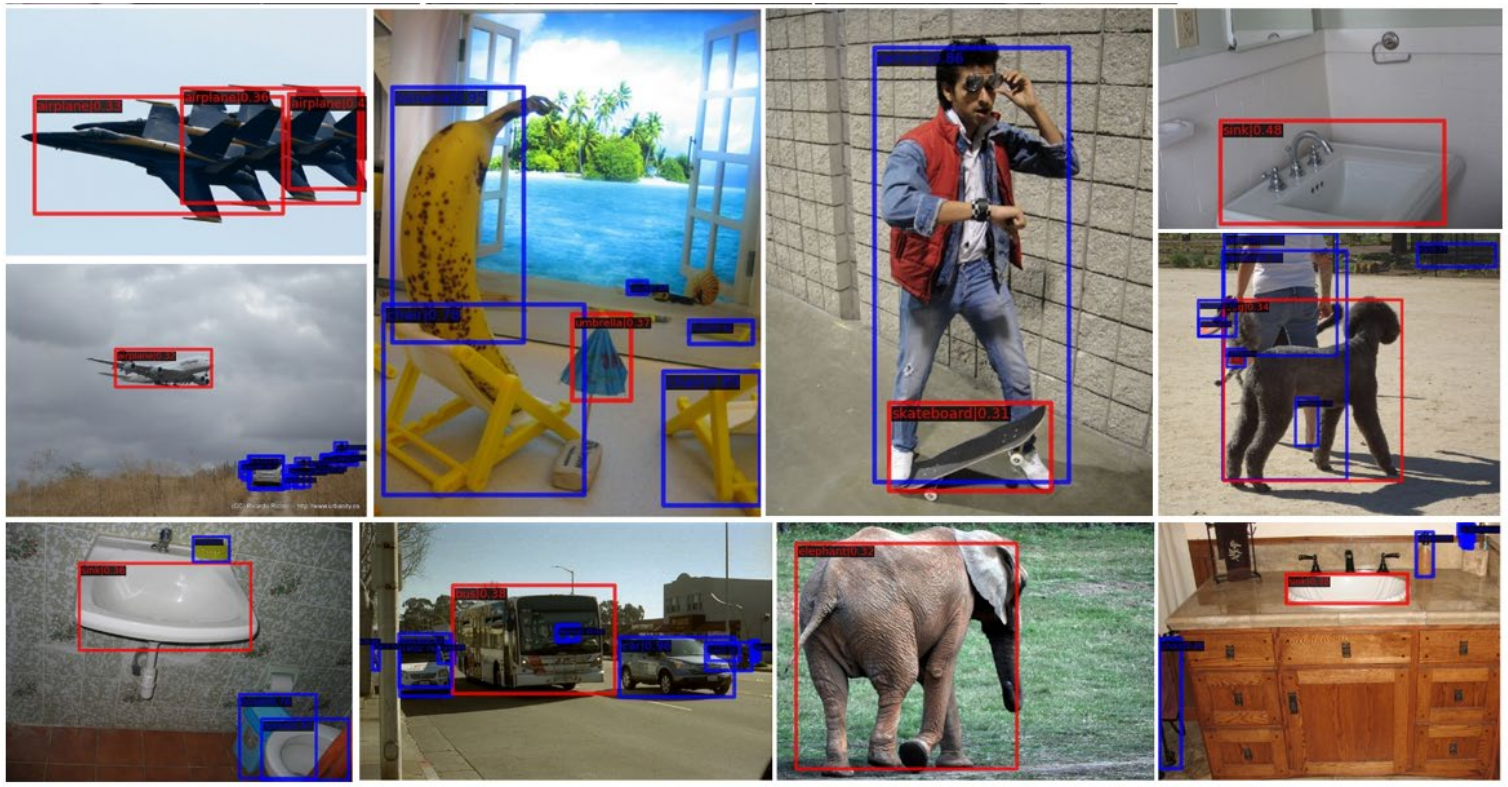



Figure A2: Visualization of object detection results. The red boxes are for the novel categories and the blue boxes are for the base categories.

# Visualization & Analysis

- Qualitative results on dense prediction tasks

Figure A3: Visualization of image segmentation. The images are from ADE20k (Zhou et al., 2017).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
S-LAB FOR ADVANCED INTELLIGENCE
商汤 sensetime
ICLR

# Thank You!