LPOSS: Label Propagation Over Patches and Pixels for Open-vocabulary Semantic Segmentation

Vladan Stojnić, Yannis Kalantidis, Jiří Matas, Giorgos Tolias

PR&CV Colloquium, April 10, 2025

Open-vocabulary semantic segmentation

- Image

Open-vocabulary semantic segmentation

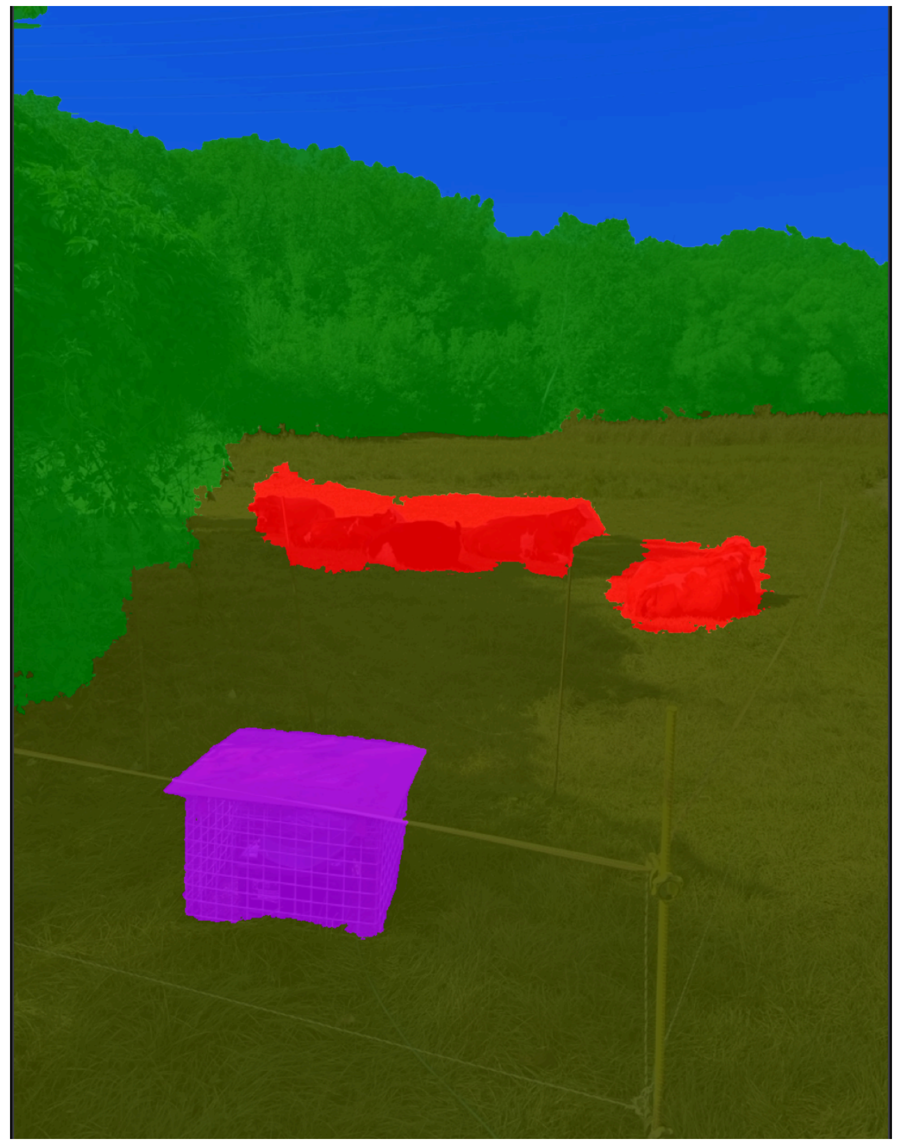
- Image

- Class names as text: **cows**, **grass**, **trees**, **sky**, **box**

Open-vocabulary semantic segmentation

- Image

- Segmentation

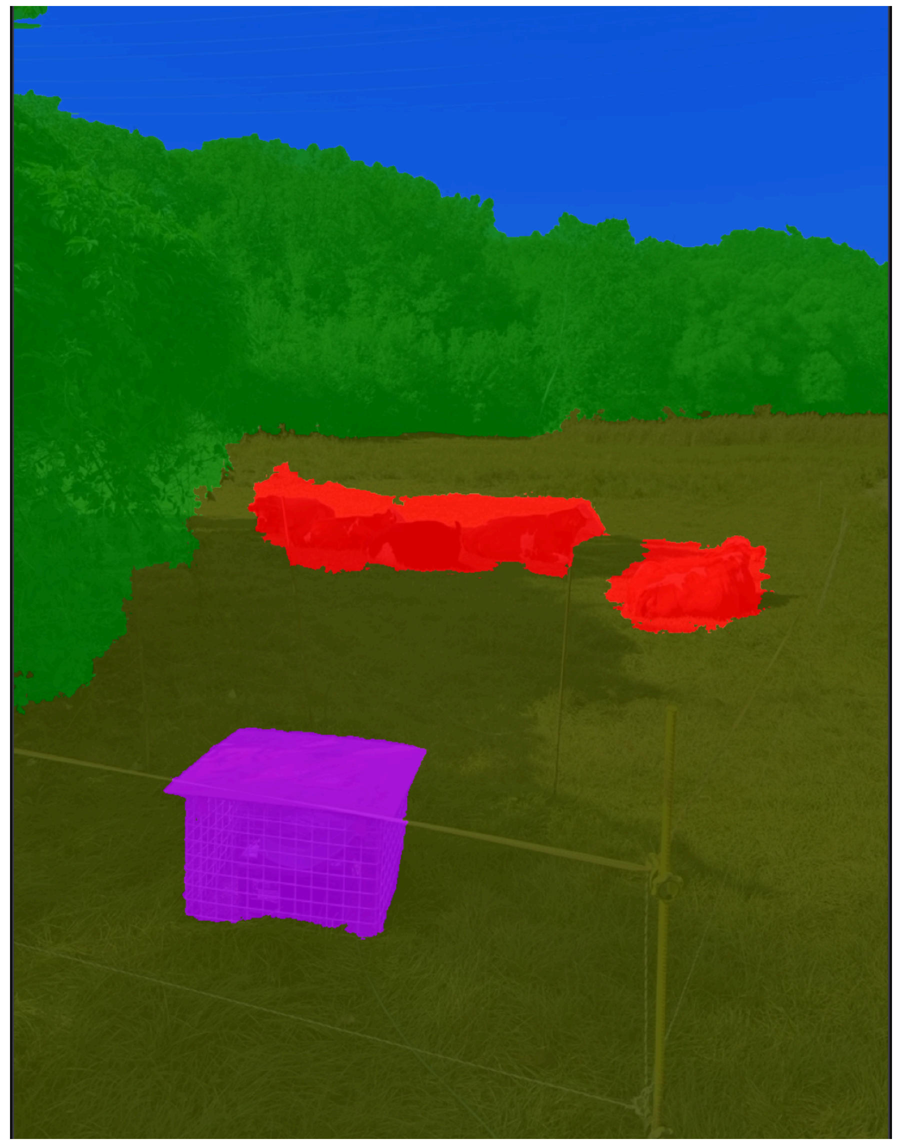


- Class names as text: **cows**, **grass**, **trees**, **sky**, **box**

Open-vocabulary semantic segmentation

- Image

- Segmentation

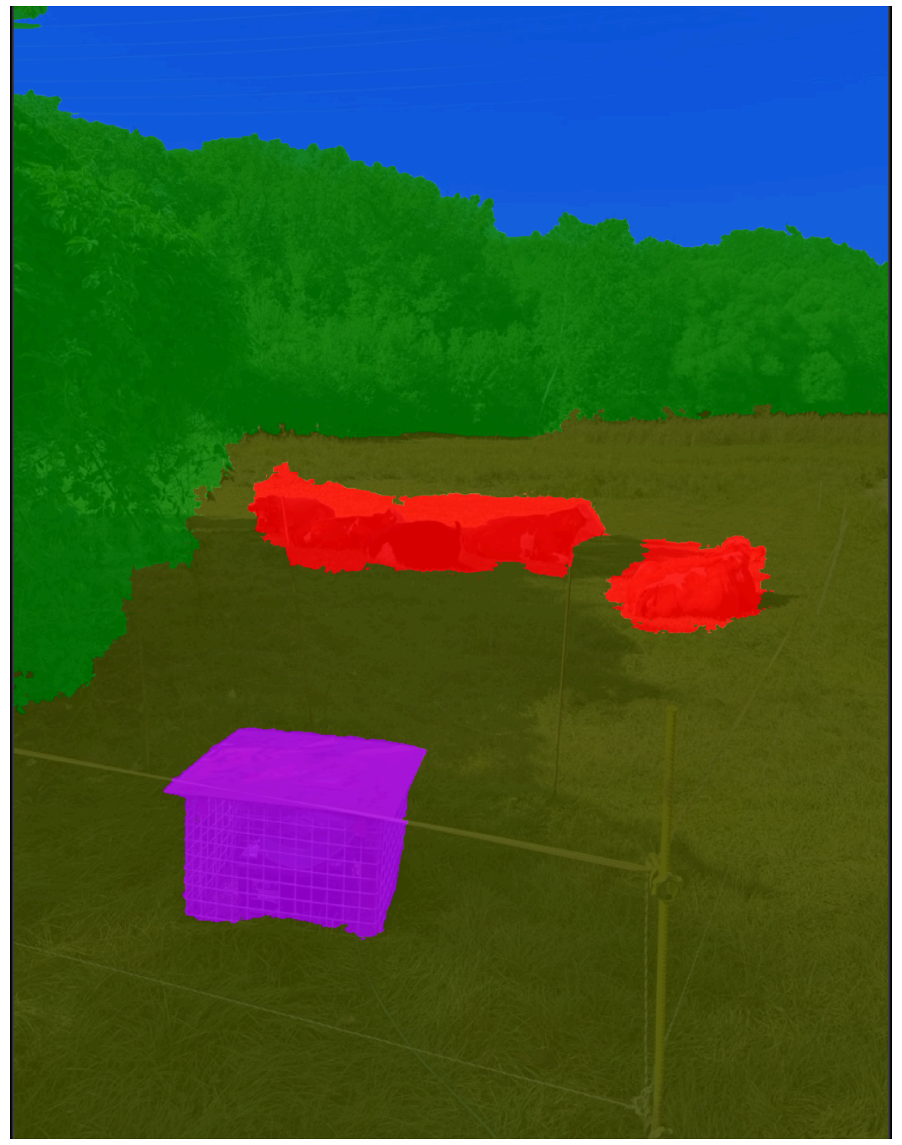


- Class names as text: cows, grass, trees, sky, box, car, people

Open-vocabulary semantic segmentation

- Image

- Segmentation



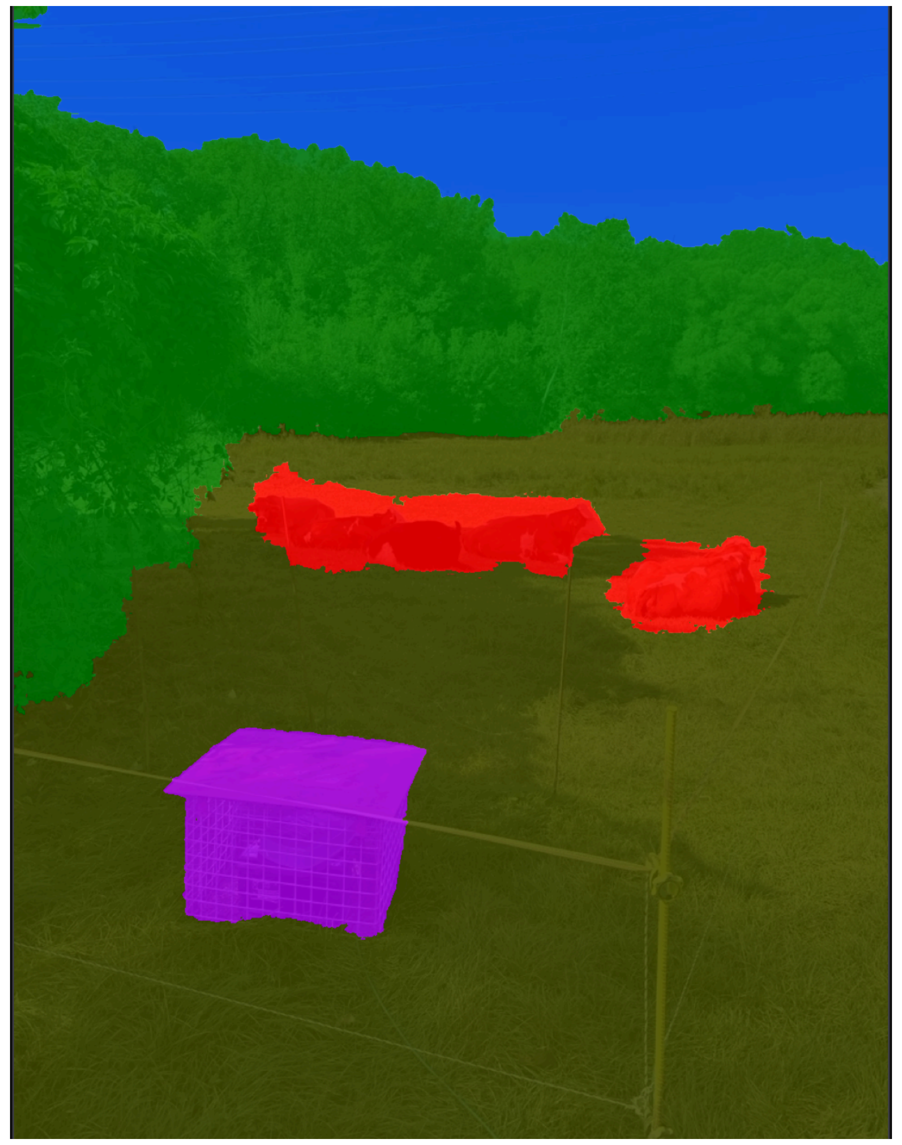
- Class names as text: cows, grass, trees, sky, box, car, people

- Open-vocabulary (zero-shot) vs open-set

Open-vocabulary semantic segmentation

- Image

- Segmentation



- Class names as text: **cows**, **grass**, **trees**, **sky**, **box**, **car**, **people** **other (or background)**

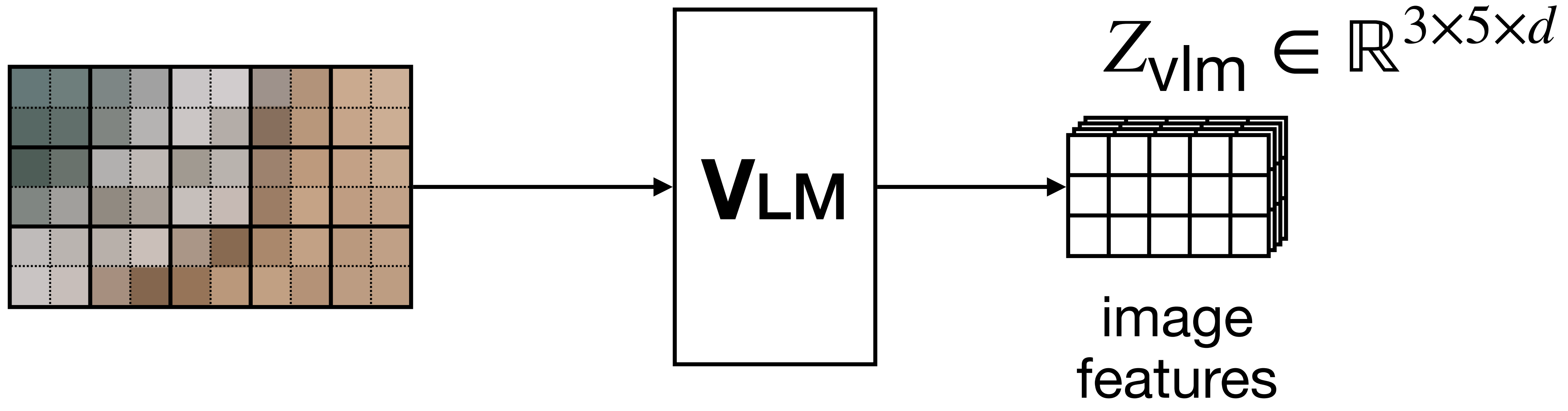
- Open-vocabulary (zero-shot) vs open-set

Use of VLMs

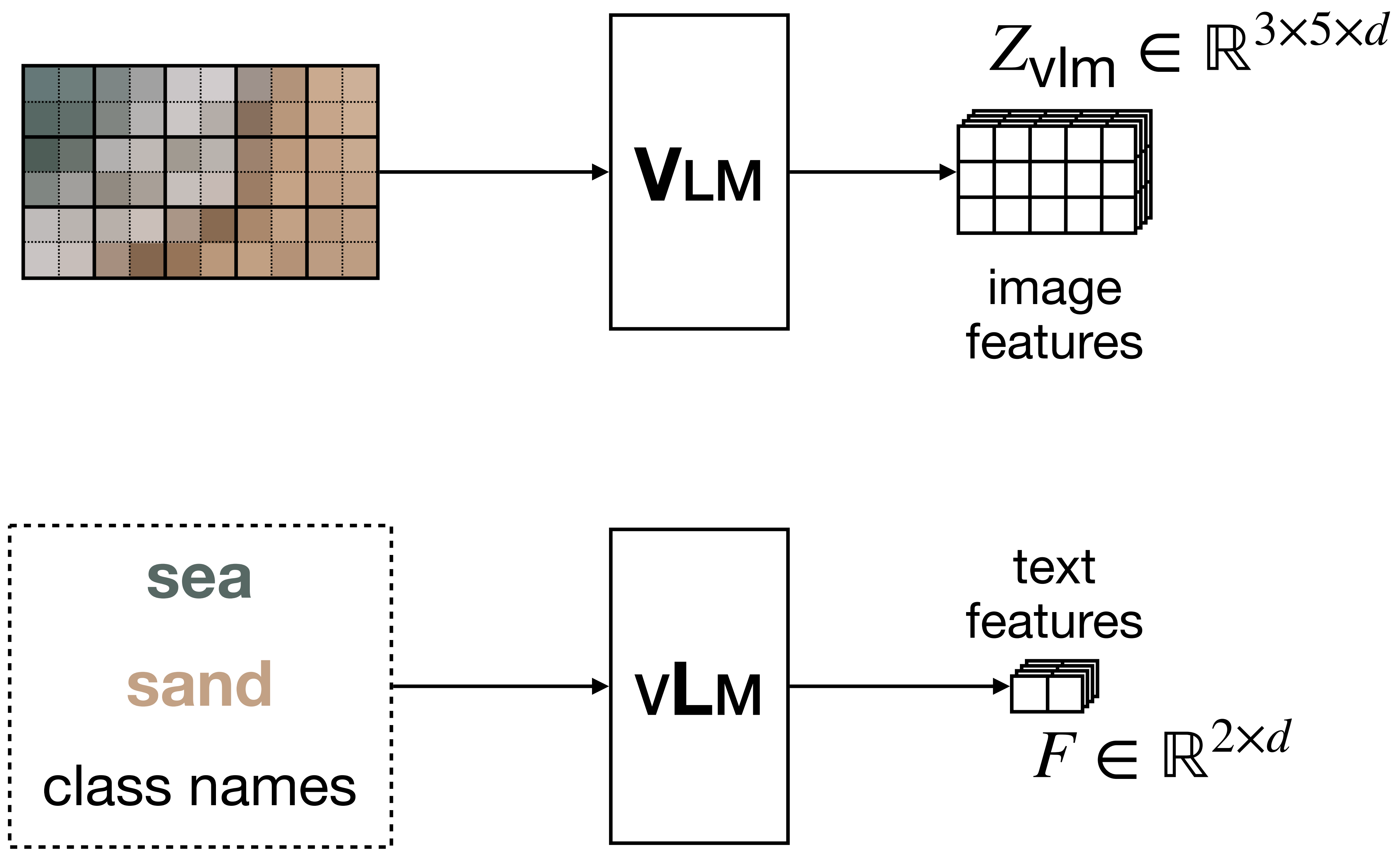
- VLMs, e.g. CLIP [1], excel in open-vocabulary tasks
  - Zero-shot classification
  - Text2image and image2text retrieval

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, et.al. Learning transferable visual models from natural language supervision. In ICML, 2021.

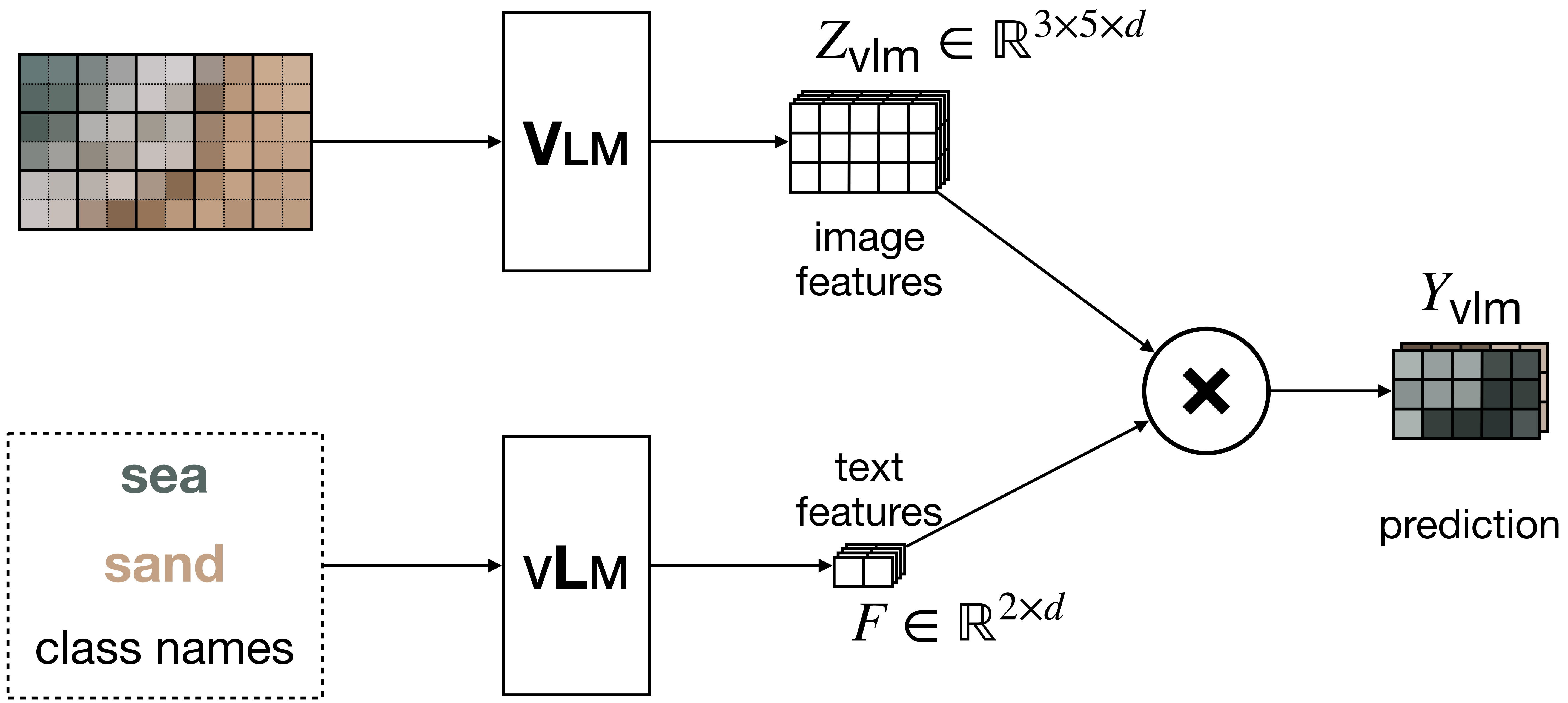
Use of VLMs for semantic segmentation



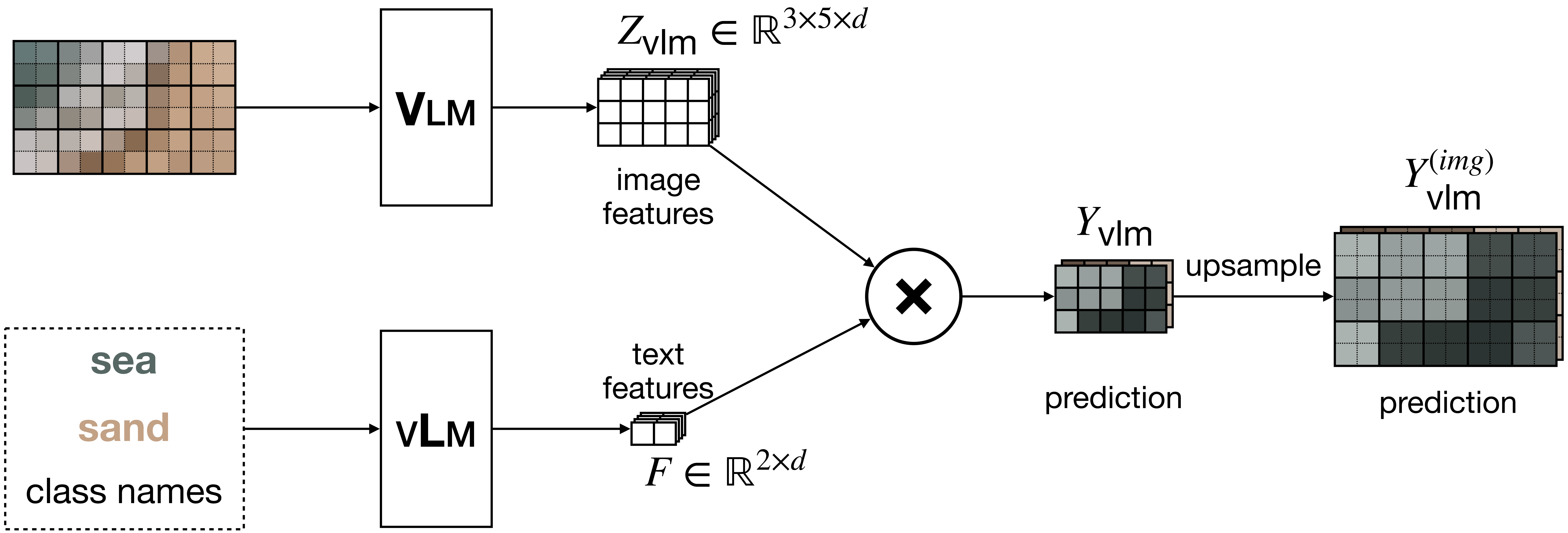
Use of VLMs for semantic segmentation



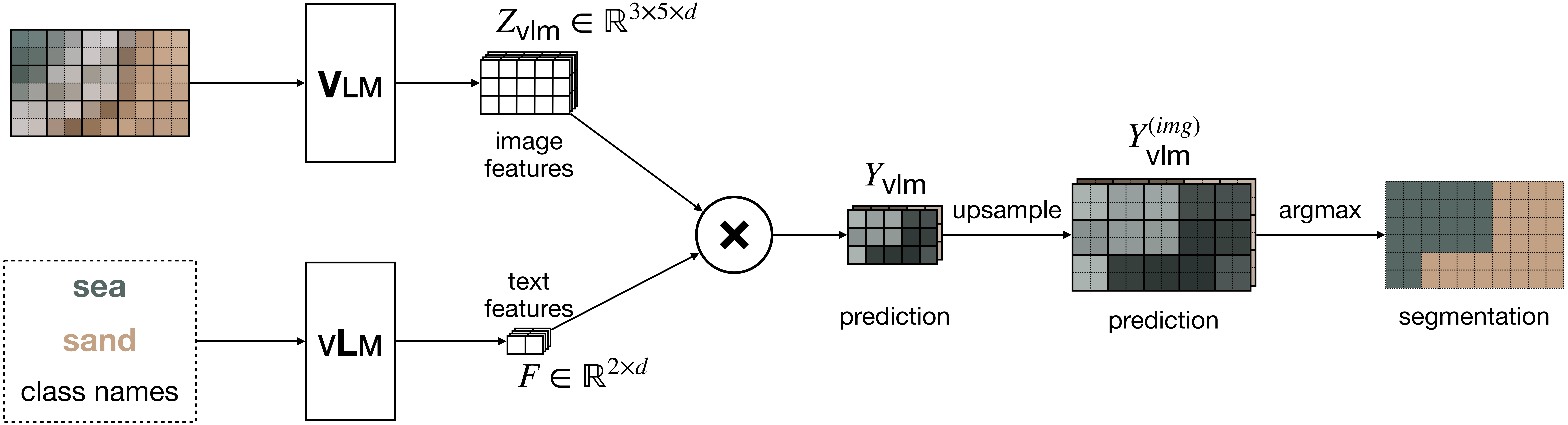
Use of VLMs for semantic segmentation



Use of VLMs for semantic segmentation



Use of VLMs for semantic segmentation



Use of VLMs for semantic segmentation

- Out of the box does not work well
  - Trained only with the global objective

[1] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In ECCV, 2022.

[2] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In CVPR, 2024.

[3] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In ECCV, 2024.

[4] Mengcheng Lan, Chaofeng Chen, Yiping Ke, et. al.. ClearCLIP: Decomposing clip representations for dense vision-language inference. In ECCV, 2024.

Use of VLMs for semantic segmentation

- Out of the box does not work well
  - Trained only with the global objective
- A lot of work on slightly modifying the ViT architecture during inference:
  - MaskCLIP [1]
  - GEM [2]
  - SCLIP [3]
  - ClearCLIP [4]

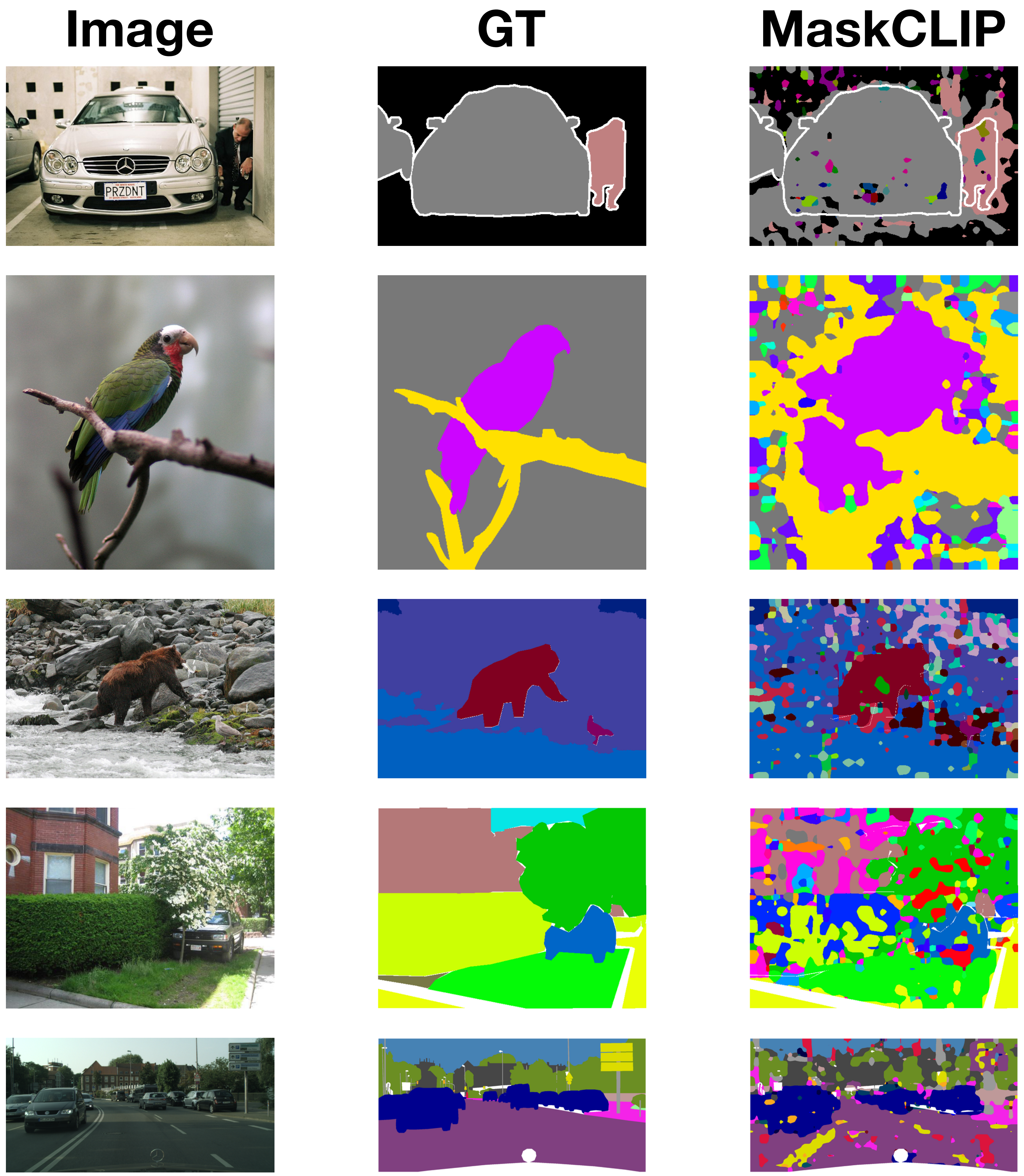
[1] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In ECCV, 2022.

[2] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In CVPR, 2024.

[3] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In ECCV, 2024.

[4] Mengcheng Lan, Chaofeng Chen, Yiping Ke, et. al.. ClearCLIP: Decomposing clip representations for dense vision-language inference. In ECCV, 2024.

Use of VLMs for semantic segmentation



mIoU: 27.0%
(average over 8 datasets)

LPOSS

- Can we improve using classical segmentation approaches?

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i

$$Q(\hat{Y}) = \sum_{i=1}^{N} f(\hat{Y}_i, Y_i)$$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches

$$Q(\hat{Y}) = \sum_{i=1}^{N} f(\hat{Y}_i, Y_i) + \sum_{(i,j)\in near} g(\hat{Y}_i, \hat{Y}_j)$$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches
- We pick $f(u, v) \sim g(u, v) = \|u - v\|^2$

$$Q(\hat{Y}) = \sum_{i=1}^{N} f(\hat{Y}_i, Y_i) + \sum_{(i,j) \in near} g(\hat{Y}_i, \hat{Y}_j)$$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches
- We pick $f(u, v) \sim g(u, v) = \|u - v\|^2$
- Label propagation solves such a problem

$$Q(\hat{Y}) = (1 - \alpha) \sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches
- We pick $f(u, v) \sim g(u, v) = \|u - v\|^2$
- Label propagation solves such a problem

adjacency matrix
- symmetric
- zero diagonal
- typically very sparse

$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches
- We pick $f(u, v) \sim g(u, v) = \|u - v\|^2$
- Label propagation solves such a problem

adjacency matrix
- symmetric
- zero diagonal
- typically very sparse

$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

degree $d_j = \sum_{k=1}^{N} S_{jk}$

LPOSS

- Can we improve using classical segmentation approaches?
  - Respect initial VLM predictions Y_i
  - Predict the same label for nearby patches
- We pick $f(u, v) \sim g(u, v) = \|u - v\|^2$
- Label propagation solves such a problem

propagation hyper-parameter

adjacency matrix
- symmetric
- zero diagonal
- typically very sparse

$$Q(\hat{Y}) = (1 - \alpha) \sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

degree $d_j = \sum_{k=1}^{N} S_{jk}$

LPOSS - adjacency S

- How to construct S?

$$S =$$

LPOSS - adjacency S

- How to construct S?
  - Appearance-based adjacency S_a
    - kNN graph based on test image patch features

$$S = S_a$$

LPOSS - adjacency S

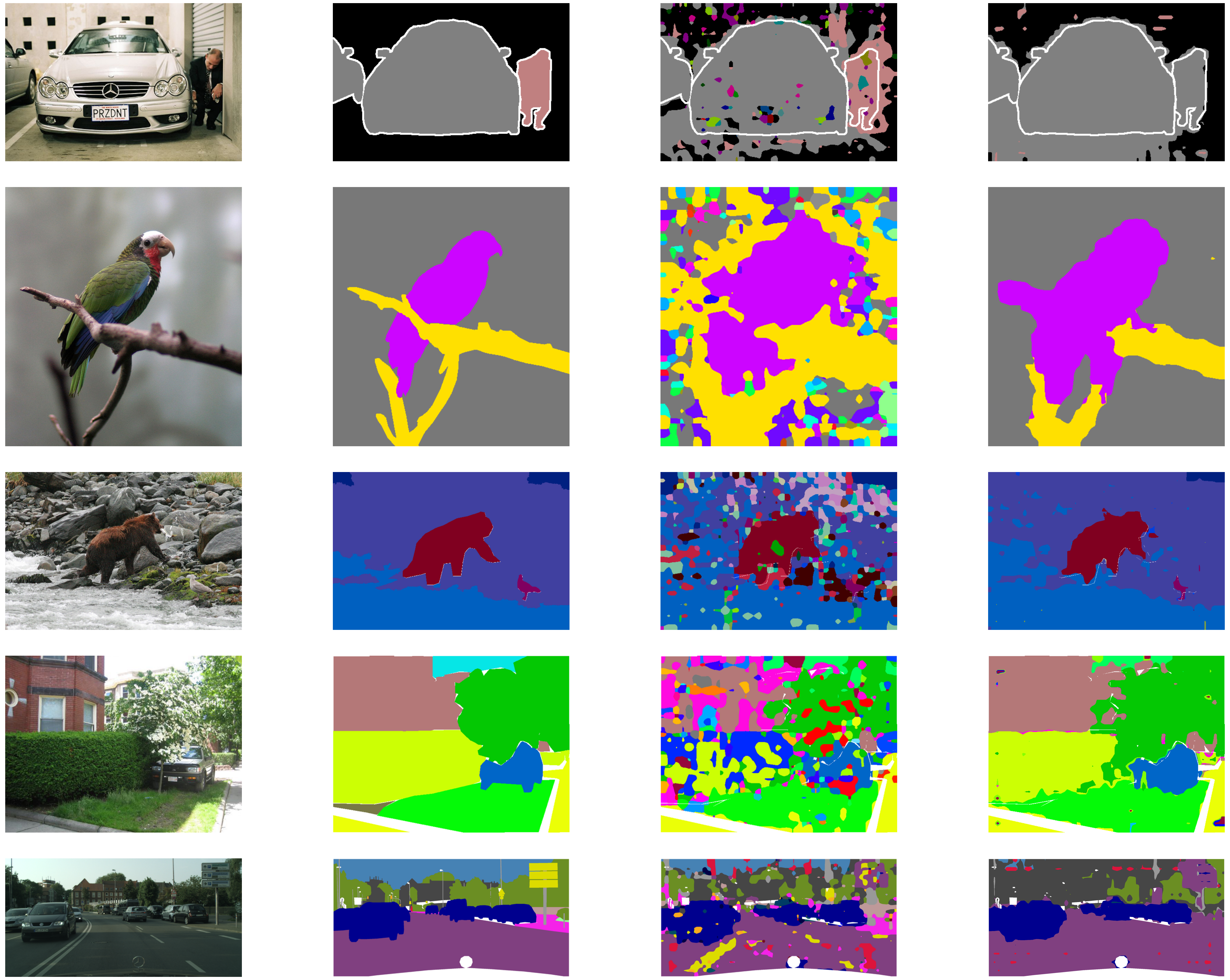
- How to construct S?
  - Appearance-based adjacency S_a
    - kNN graph based on test image patch features
  - Spatial-based adjacency S_p
    - Depends on the distance between patches

$$S = S_a \odot S_p$$

Hadamard product

LPOSS

Image | **GT** | **MaskCLIP** | **LPOSS**



mIoU: 27.0% mIoU: 38.3%

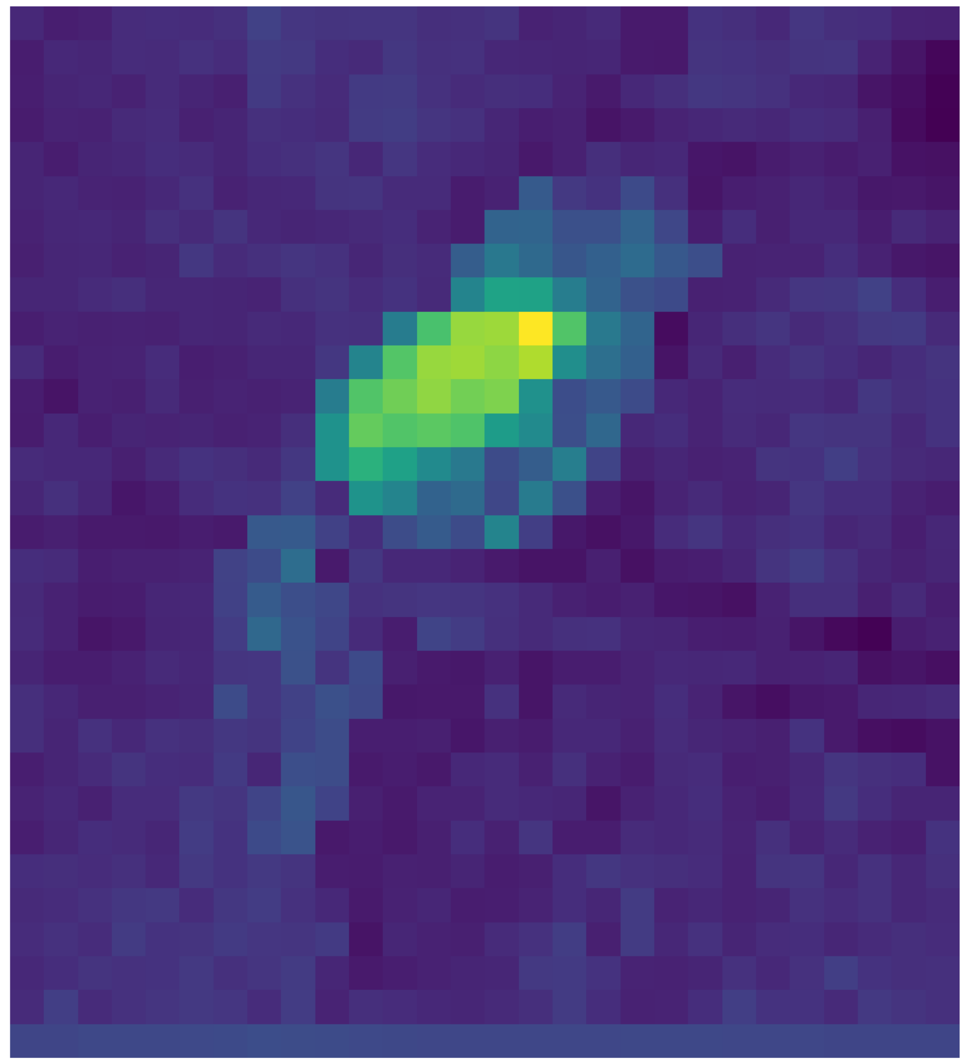
(average over 8 datasets)

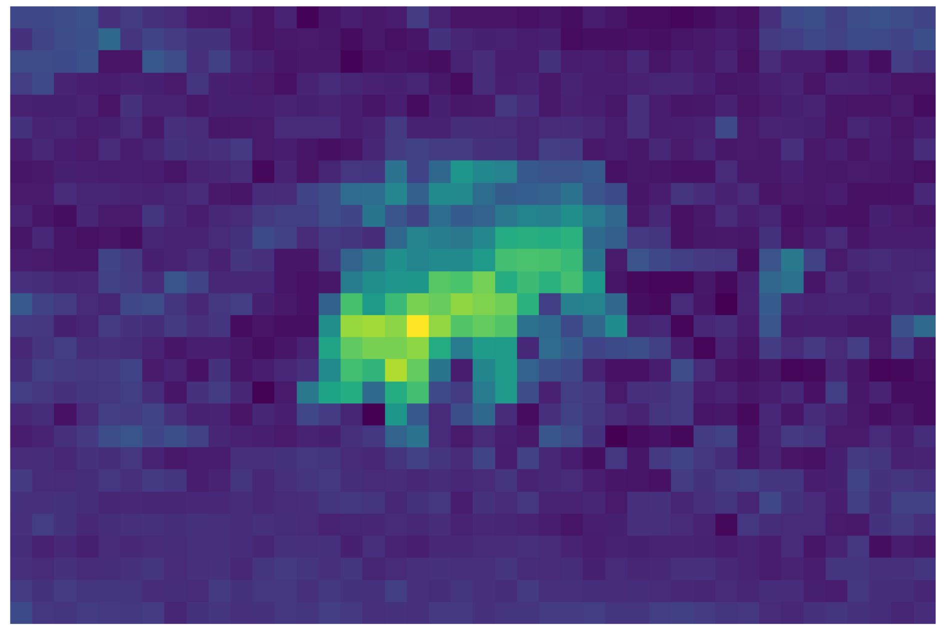
LPOSS - adjacency S

- appearance-based adjacency S_a is based on VLM features

LPOSS - adjacency S

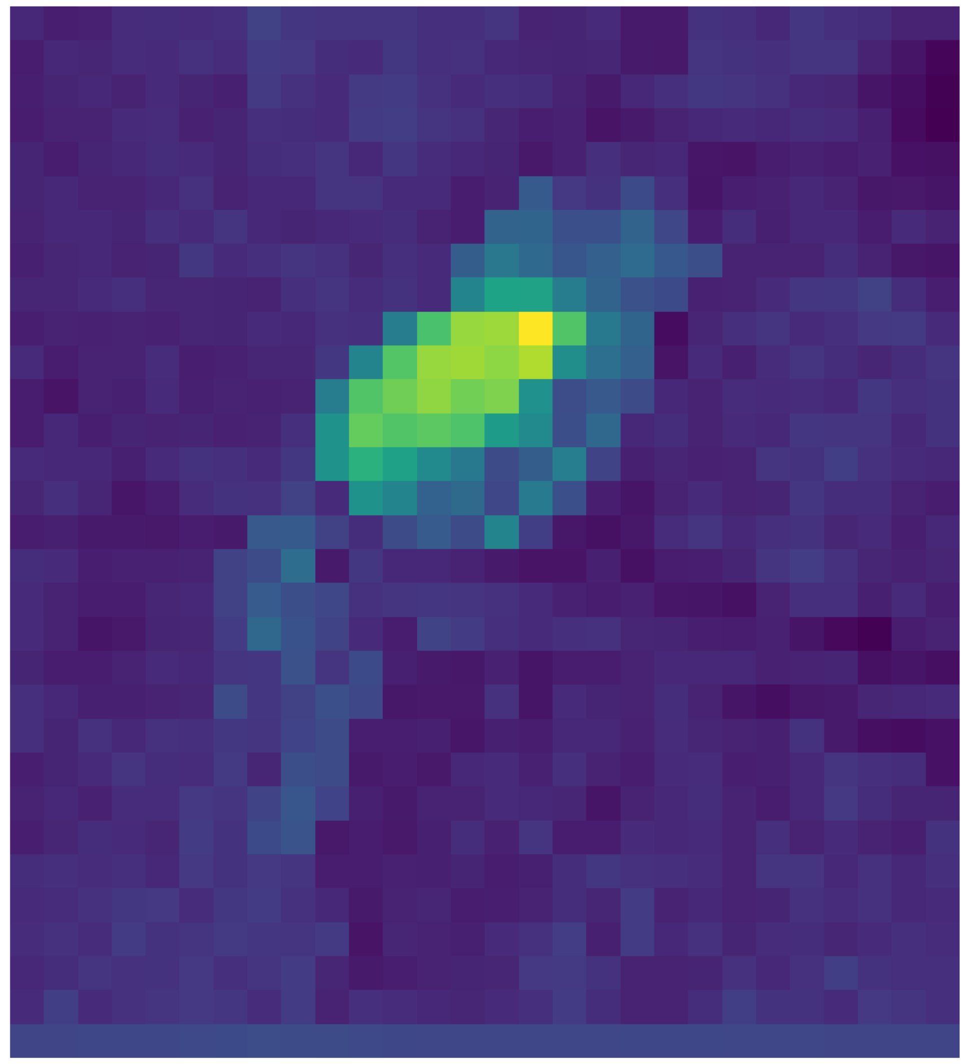
- appearance-based adjacency S_a is based on VLM features
- SSL vision models (VMs), e.g. DINO, have good localization properties

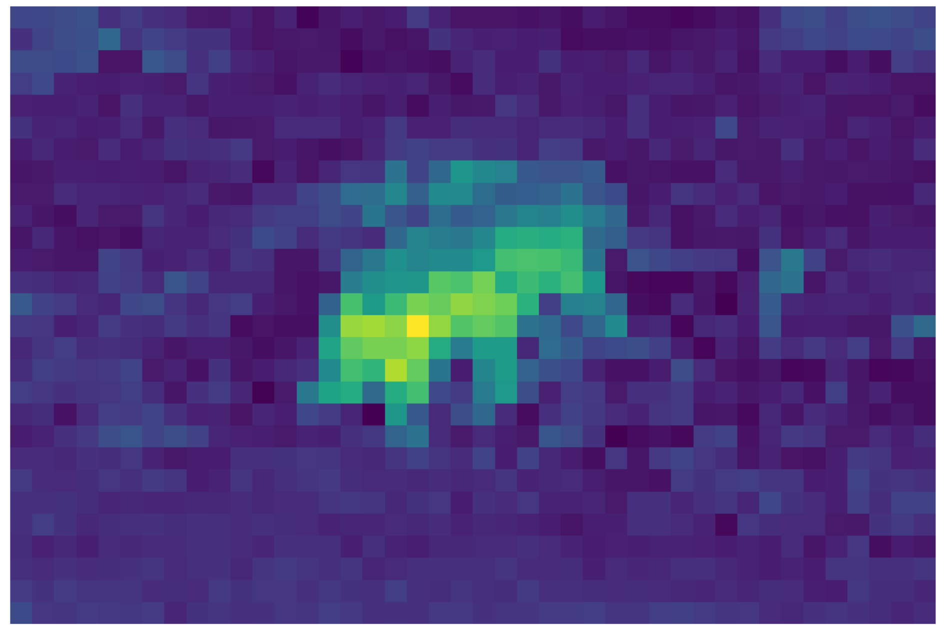





LPOSS - adjacency S

- appearance-based adjacency S_a is based on VLM features
- SSL vision models (VMs), e.g. DINO, have good localization properties



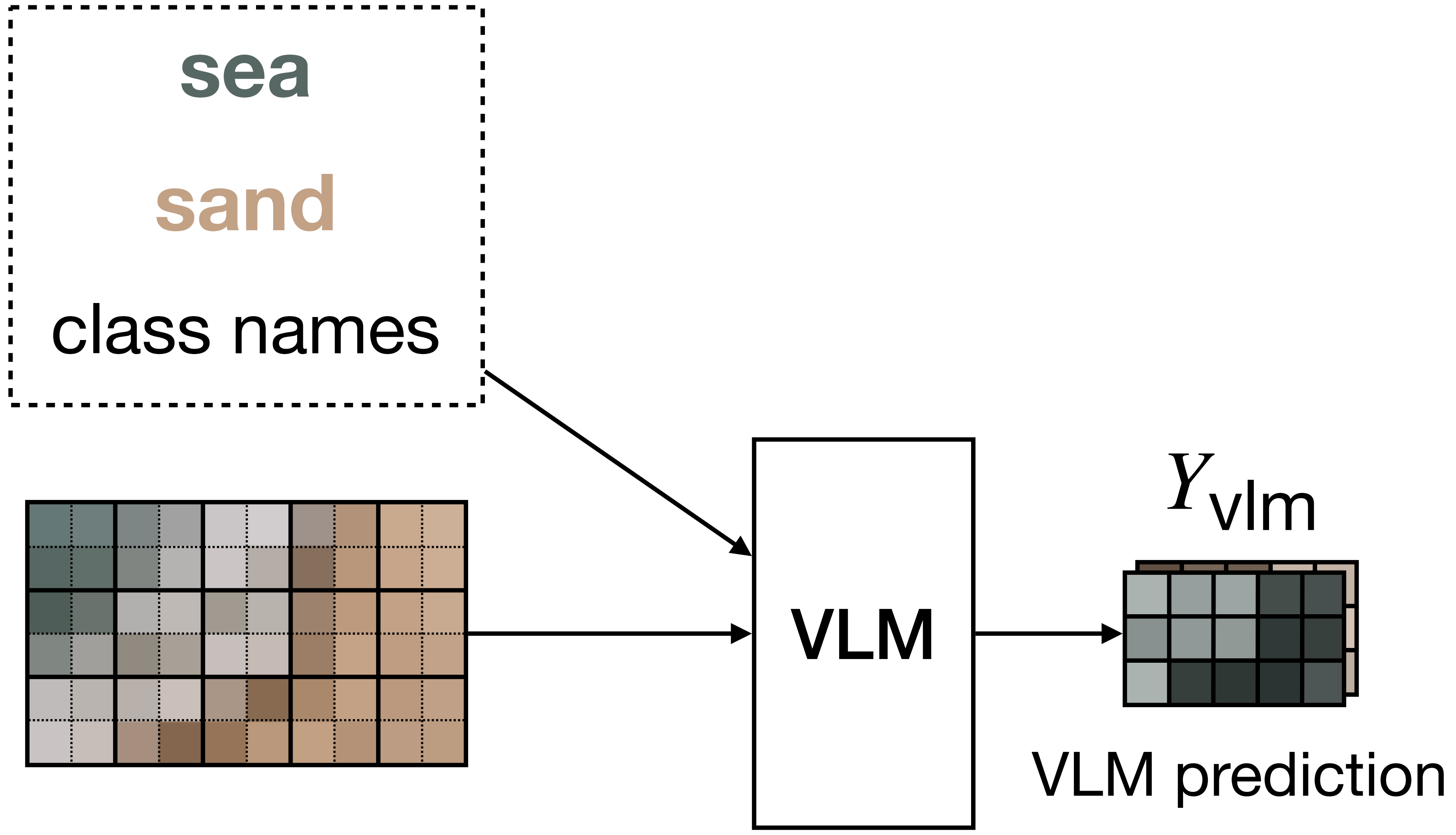
- Use VM features for appearance-based adjacency S_a

[1] Monika Wysoczanska, Oriane Simeoni, Michael Ramamonjisoa, et.al. CLIP-DINOiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In ECCV, 2024.

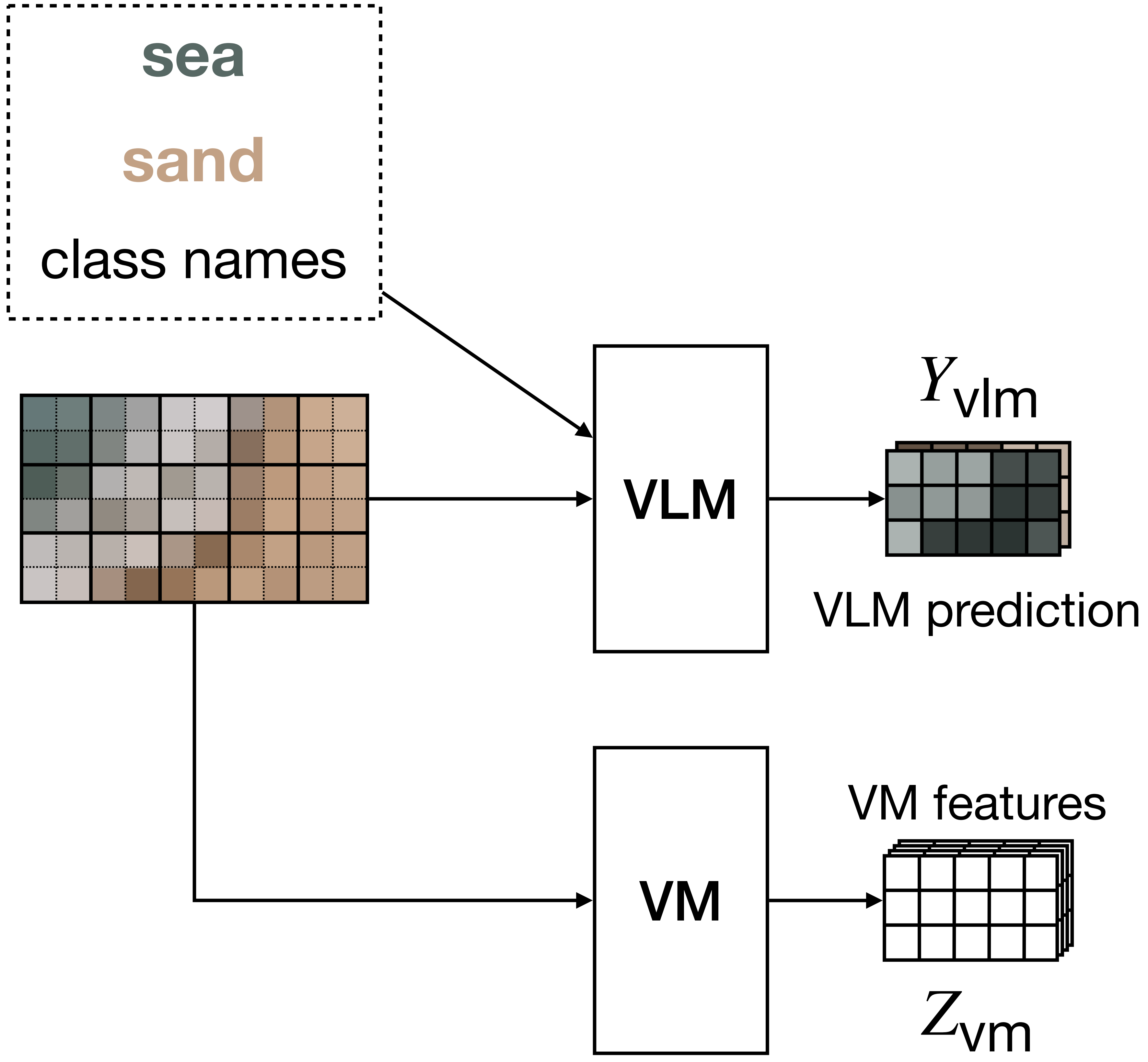
[2] Mengcheng Lan, Chaofeng Chen, Yiping Ke, et.al. ProxyCLIP: Proxy attention improves clip for open-vocabulary segmentation. In ECCV, 2024.

[3] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In ECCV, 2024.

LPOSS

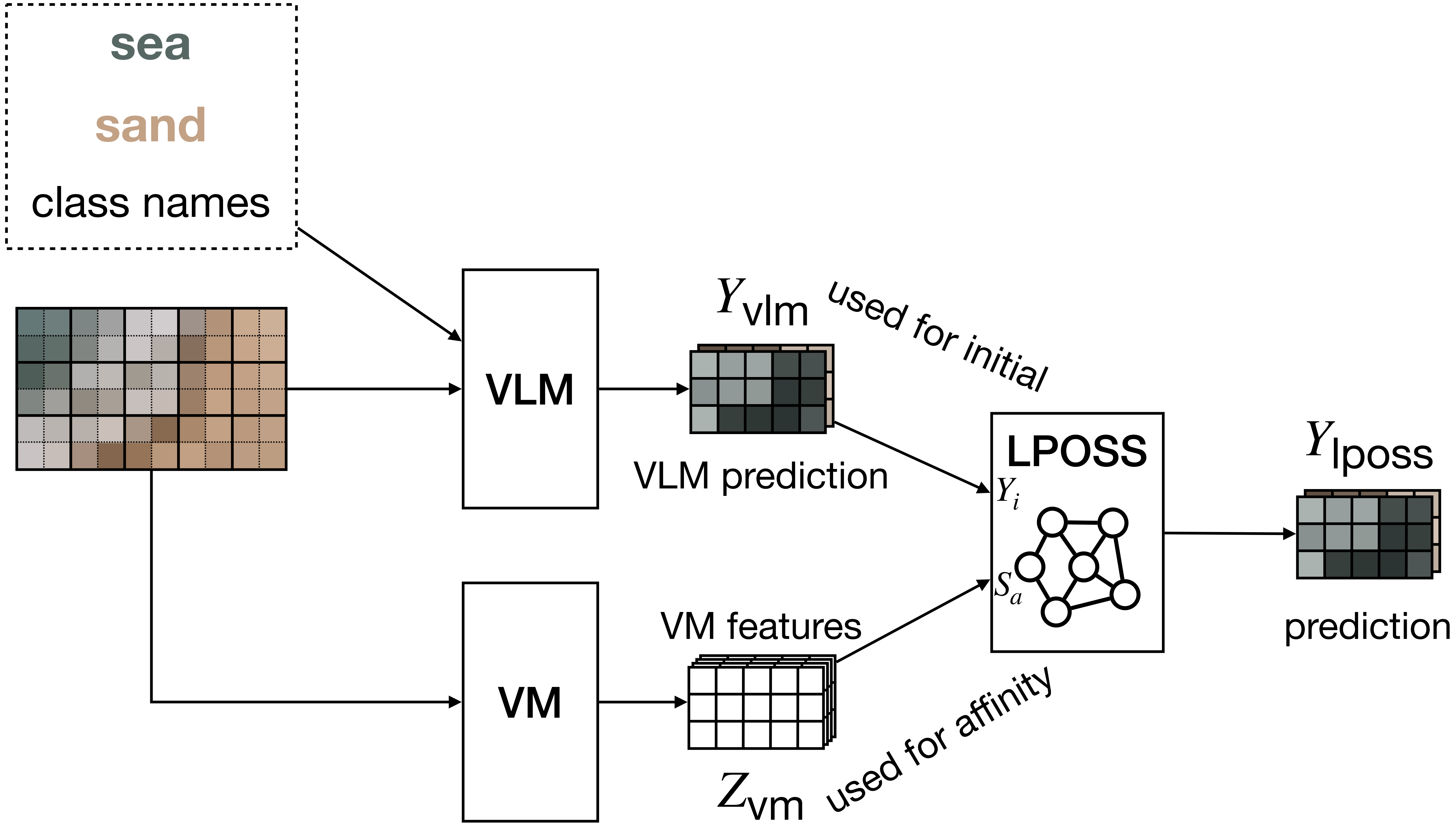


LPOSS



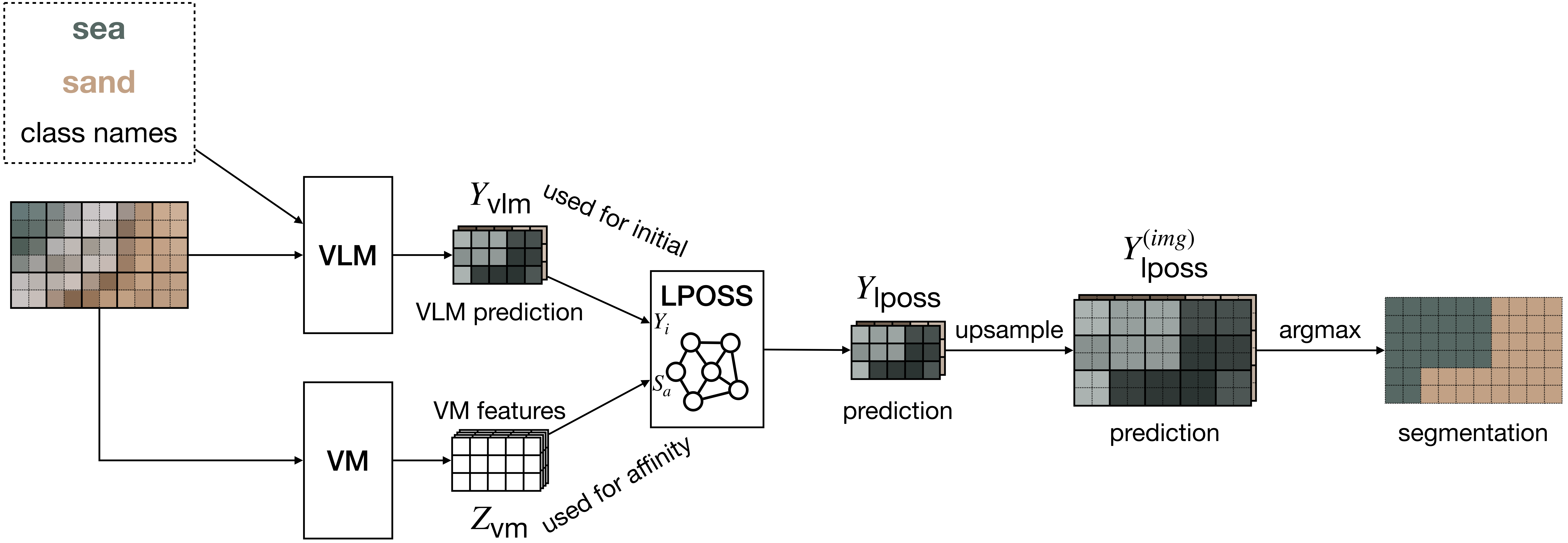
sea
sand
class names
VLM
Y_{vlm}
VLM prediction
VM
VM features
Z_{vm}

LPOSS



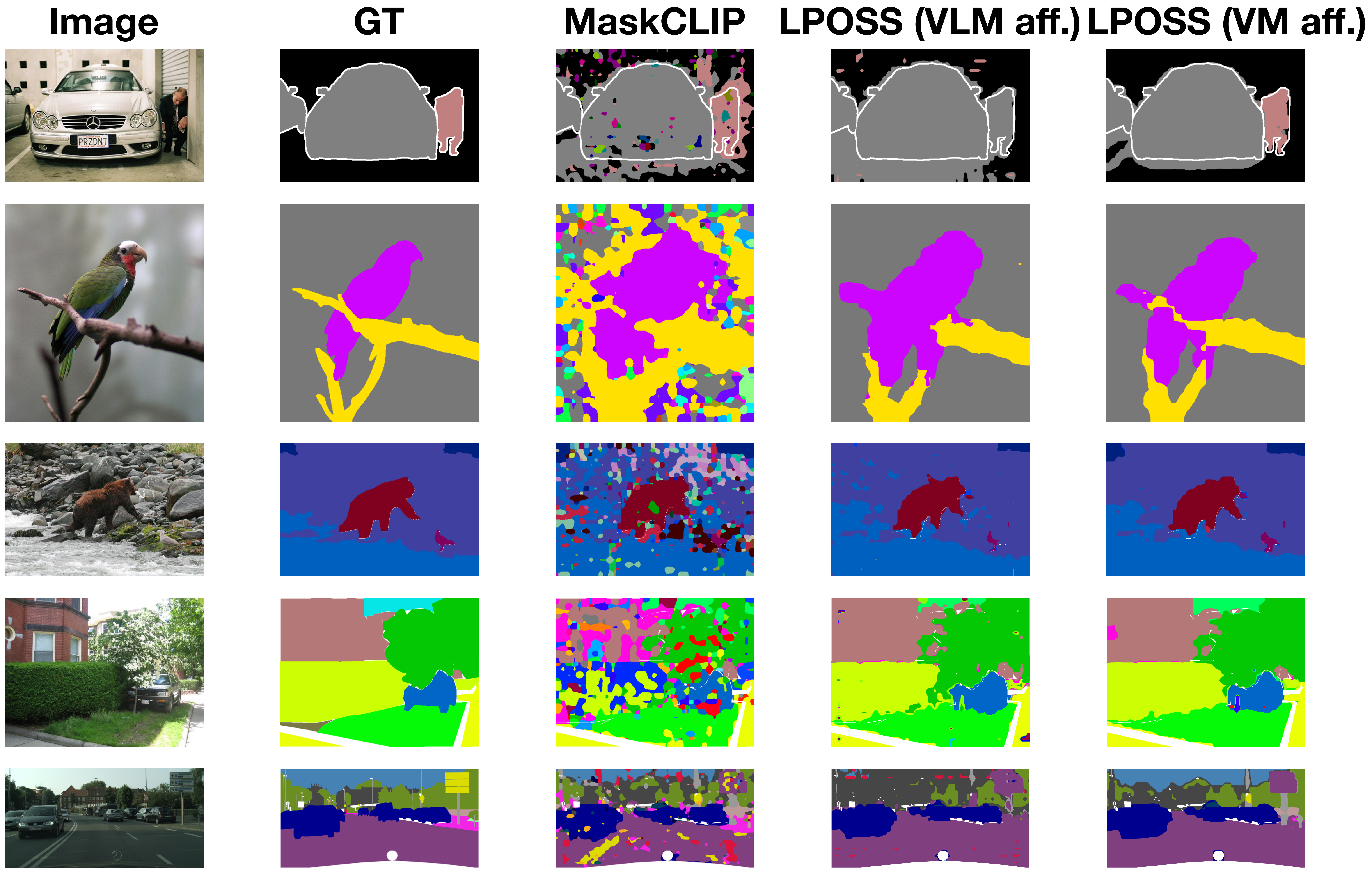
sea
sand
class names
VLM
VM
Y_{vlm}
used for initial
VLM prediction
VM features
Z_{vm}
used for affinity
LPOSS
Y_i
S_a
Y_{lposs}
prediction

LPOSS



sea
sand
class names
VLM
VM
Y_{vlm}
used for initial
VLM prediction
VM features
Z_{vm}
used for affinity
LPOSS
Y_i
S_a
Y_{lposs}
prediction
upsample
$Y_{\text{lposs}}^{(img)}$
prediction
argmax
segmentation

LPOSS



mIoU: 27.0% mIoU: 38.3% mIoU: 41.3%

(average over 8 datasets)

Relation to earlier work

time

Energy minimization
MRFs, CRFs, etc.

Classic semantic segmentation
Train on a closed set of classes with labeled images

Q_{en}^{seg}

Relation to earlier work

time

Energy minimization
MRFs, CRFs, etc.

CNN-based approaches

Classic semantic segmentation
Train on a closed set of classes with labeled images

$$Q_{en}^{seg} \ll Q_{cnn}^{seg}$$

Relation to earlier work

time

	Energy minimization MRFs, CRFs, etc.		CNN-based approaches		ViT-based approaches
Classic semantic segmentation Train on a closed set of classes with labeled images	Q^{seg}_{en}	$\ll$	Q^{seg}_{cnn}	$\ll$	Q^{seg}_{vit}

Relation to earlier work

time

	Energy minimization MRFs, CRFs, etc.		CNN-based approaches		ViT-based approaches
Classic semantic segmentation Train on a closed set of classes with labeled images	Q_{en}^{seg}	$\ll$	Q_{cnn}^{seg}	$\ll$	Q_{vit}^{seg}
					$\gg$
Open-vocabulary semantic segmentation Zero-shot evaluation on class names given as text					Q_{vlm}^{ovs}

Relation to earlier work

time

	Energy minimization MRFs, CRFs, etc.		CNN-based approaches		ViT-based approaches
Classic semantic segmentation Train on a closed set of classes with labeled images	Q^{seg}_{en}	$\ll$	Q^{seg}_{cnn}	$\ll$	Q^{seg}_{vit}
					$\Downarrow\!\gg$
					Q^{ovs}_{vlm}
Open-vocabulary semantic segmentation Zero-shot evaluation on class names given as text					$\Uparrow\!\gg$
					Q^{ovs}_{lposs}

Limitation of patch level prediction

- Predictions are on the level of patches

Limitation of patch level prediction

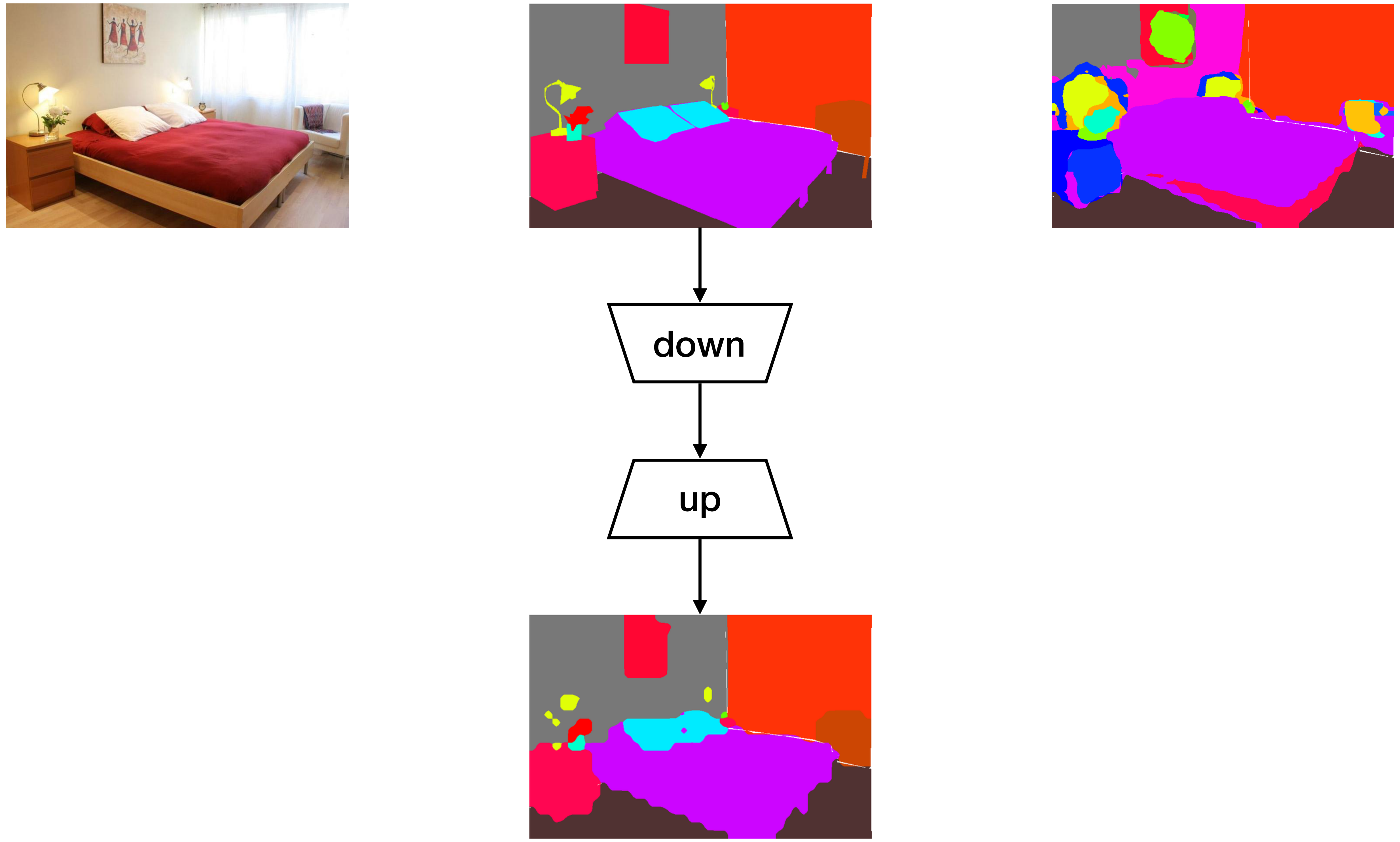
- Predictions are on the level of patches

Image | **GT** | **LPOSS**

Limitation of patch level prediction

- Predictions are on the level of patches

Image **GT** **LPOSS**



mIoU: 85.2%
Boundary IoU [1]: 69.5%
(average over 8 datasets)

[1] Bowen Cheng, Ross Girshick, Piotr Dollar, et.al. Boundary IoU: Improving object-centric image segmentation evaluation. In CVPR, 2021.

LPOSS+

- Predictions are on the level of patches

Image

GT

LPOSS

- Apply another label propagation to refine predictions on the pixel level

$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

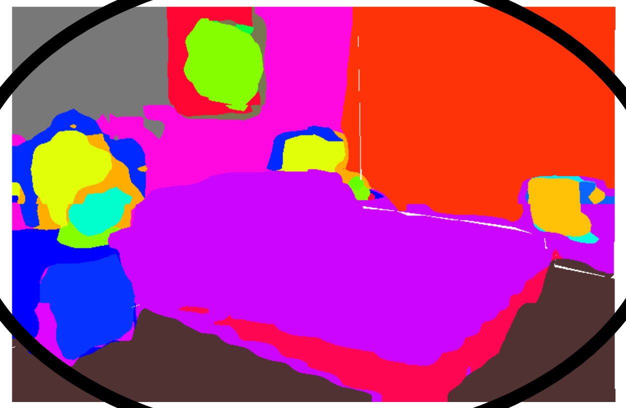
LPOSS+

- Predictions are on the level of patches

Image

GT

LPOSS



- Apply another label propagation to refine predictions on the pixel level

initialize using LPOSS predictions

$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

LPOSS+

- Predictions are on the level of patches

Image

GT

LPOSS

- Apply another label propagation to refine predictions on the pixel level

affinity over pixels
- appearance using color based features
- spatial

initialize using LPOSS predictions

$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

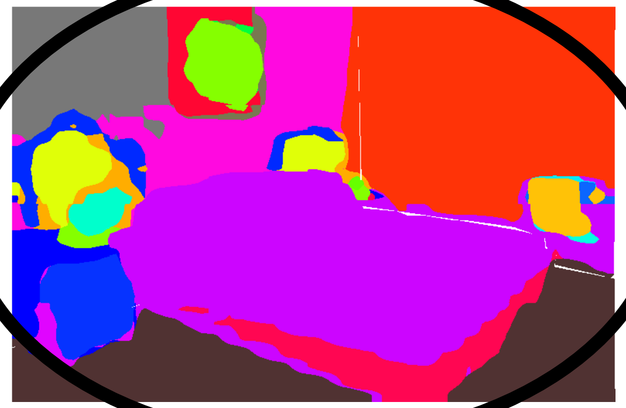
LPOSS+

- Predictions are on the level of patches

Image

GT

LPOSS



- Apply another label propagation to refine predictions on the pixel level

affinity over pixels
- appearance using color based features
- spatial

initialize using LPOSS predictions

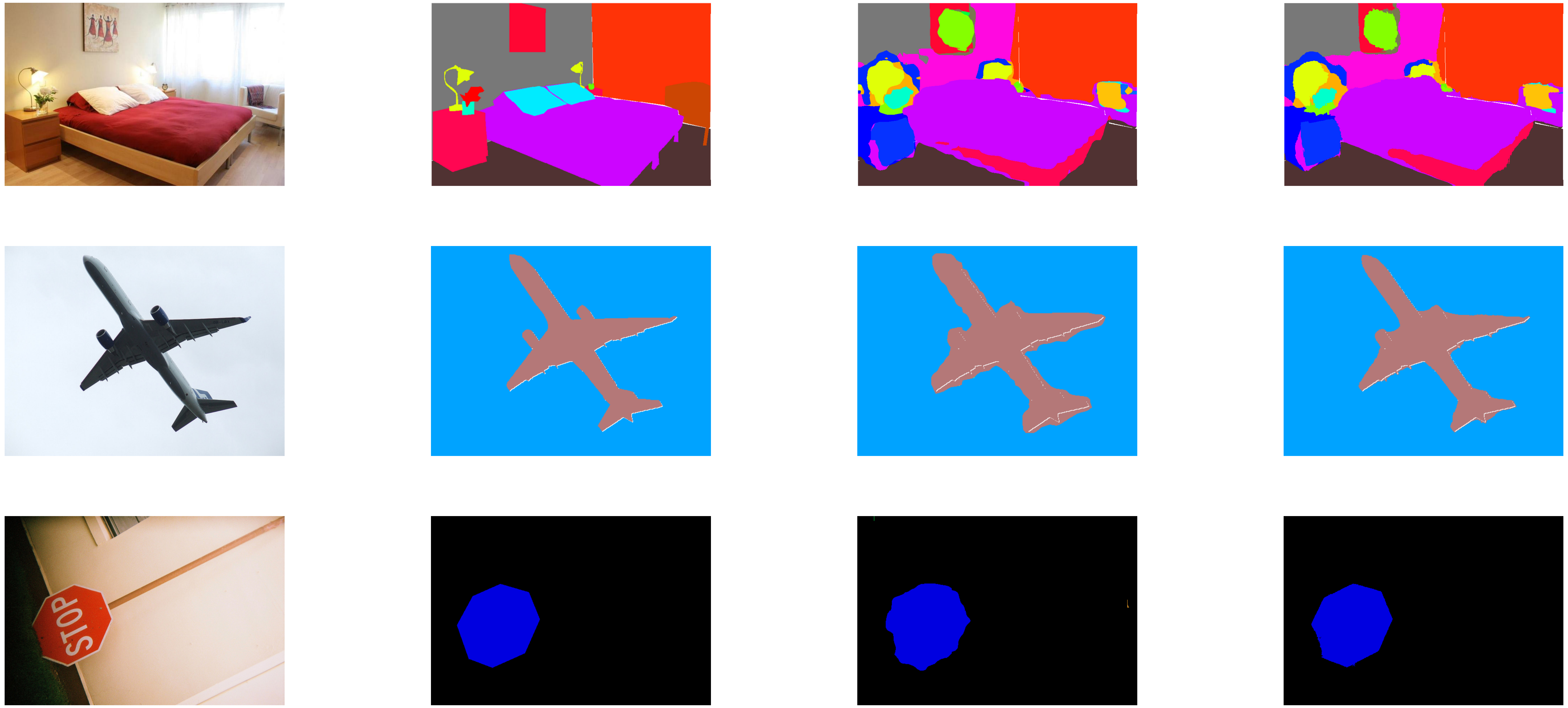
$$Q(\hat{Y}) = (1-\alpha)\sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|^2 + \alpha \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\hat{Y}_i}{\sqrt{d_i}} - \frac{\hat{Y}_j}{\sqrt{d_j}} \right\|^2$$

sum over pixels

LPOSS+

- Predictions are on the level of patches

Image **GT** **LPOSS** **LPOSS+**



mIoU: 41.3%
Boundary IoU: 30.3%

mIoU: 42.1%
Boundary IoU: 32.1%

(average over 8 datasets)

Sliding window inference

- Models trained with fixed squared resolution

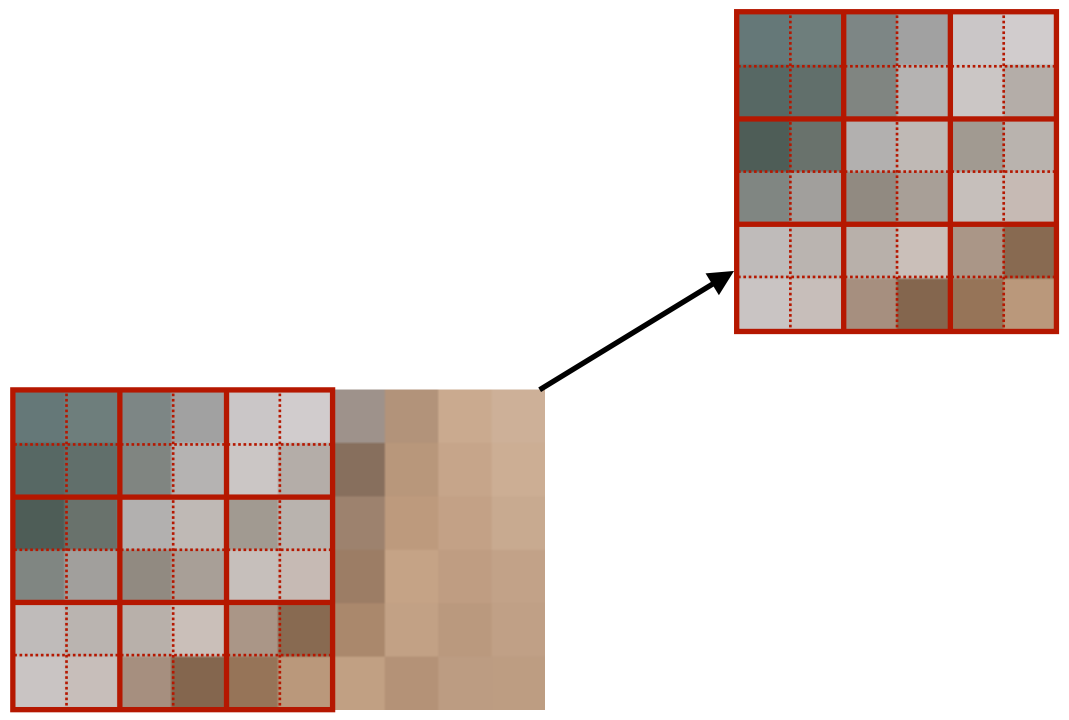
Sliding window inference

- Models trained with fixed squared resolution
- During inference
  - Different aspect ratio

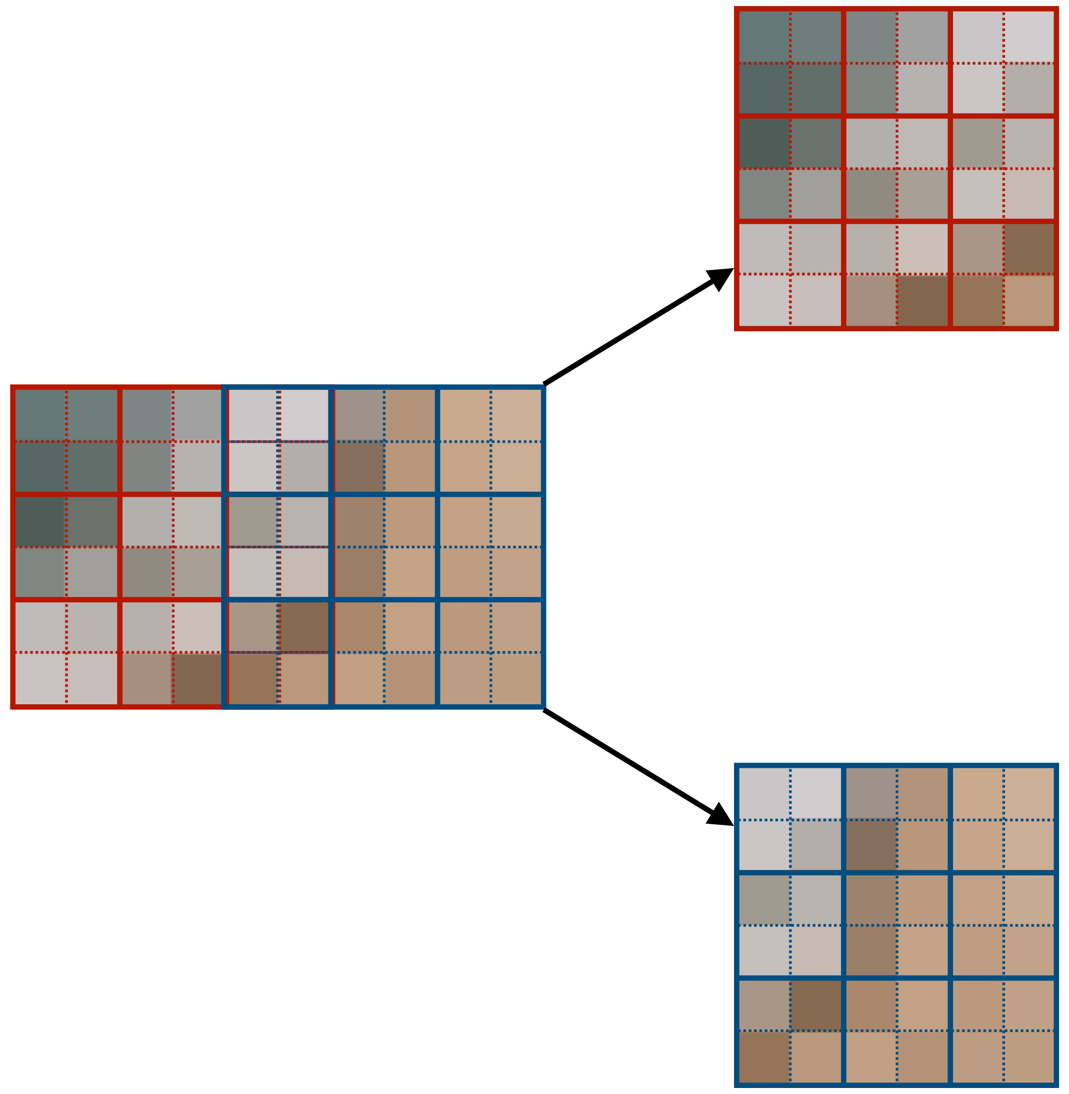
Sliding window inference

- Models trained with fixed squared resolution
- During inference
    - Different aspect ratio
    - Different resolution - different number of tokens

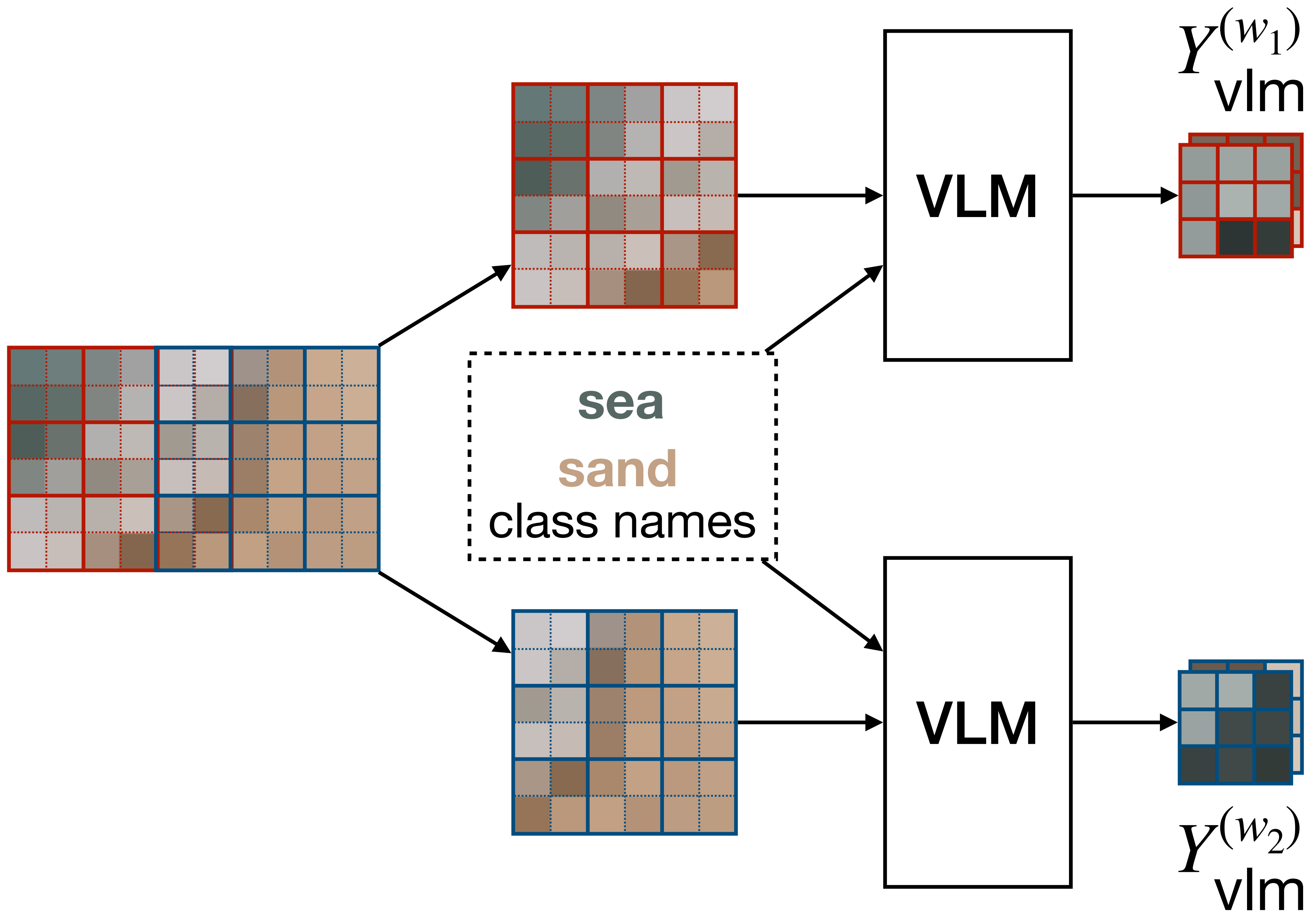
Sliding window inference



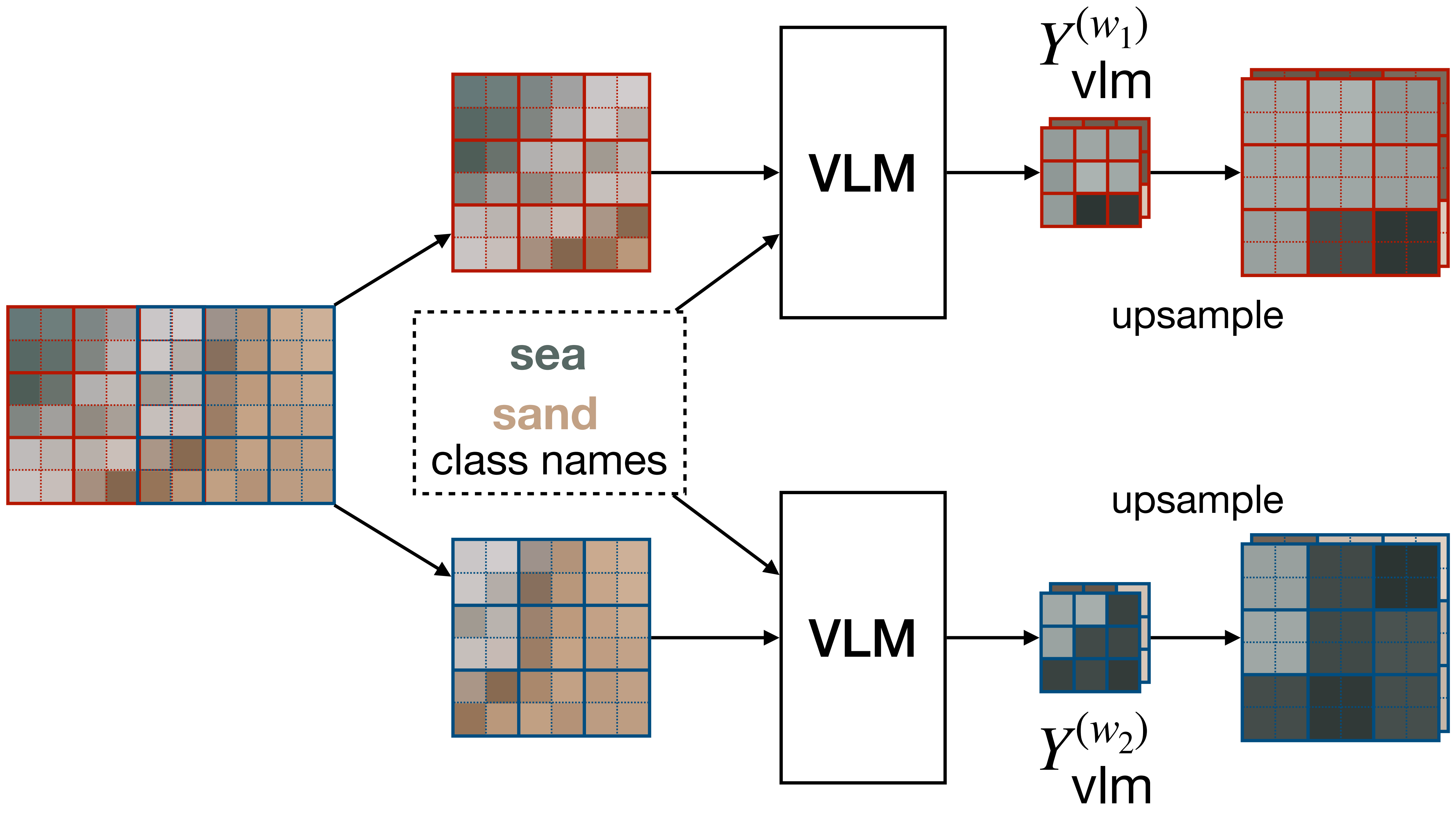
Sliding window inference



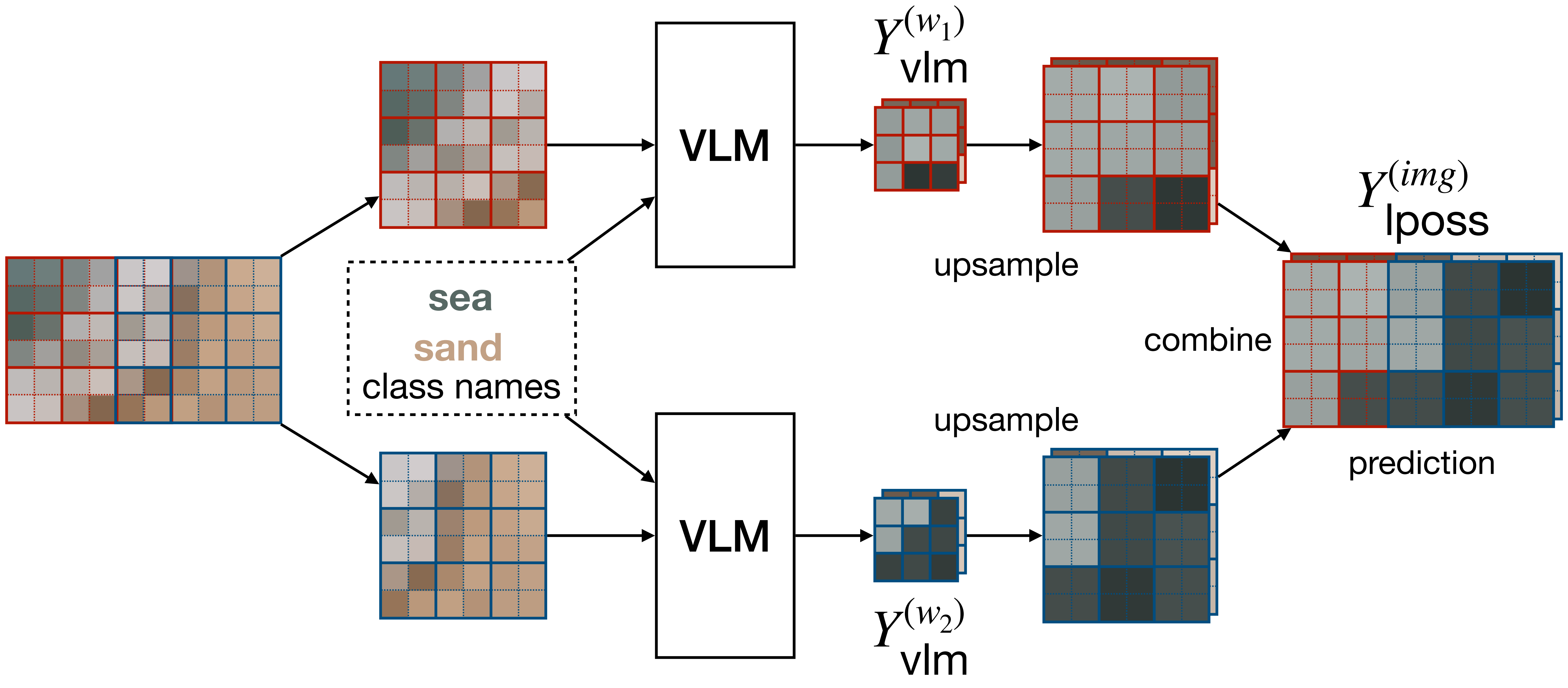
Sliding window inference



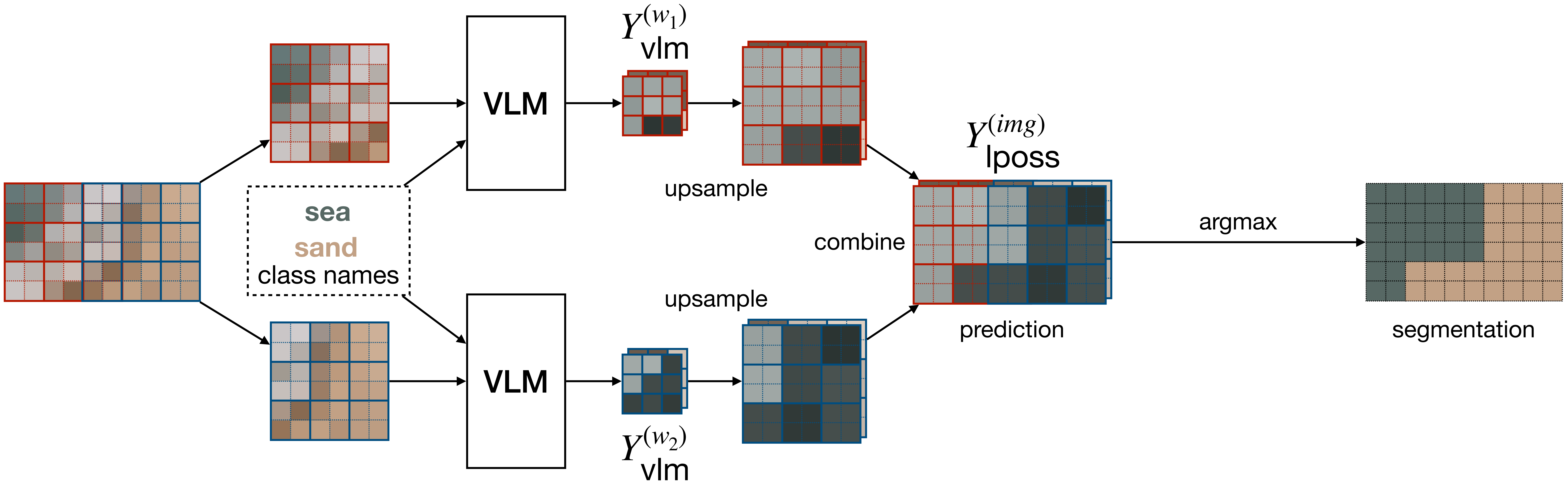
Sliding window inference



Sliding window inference



Sliding window inference



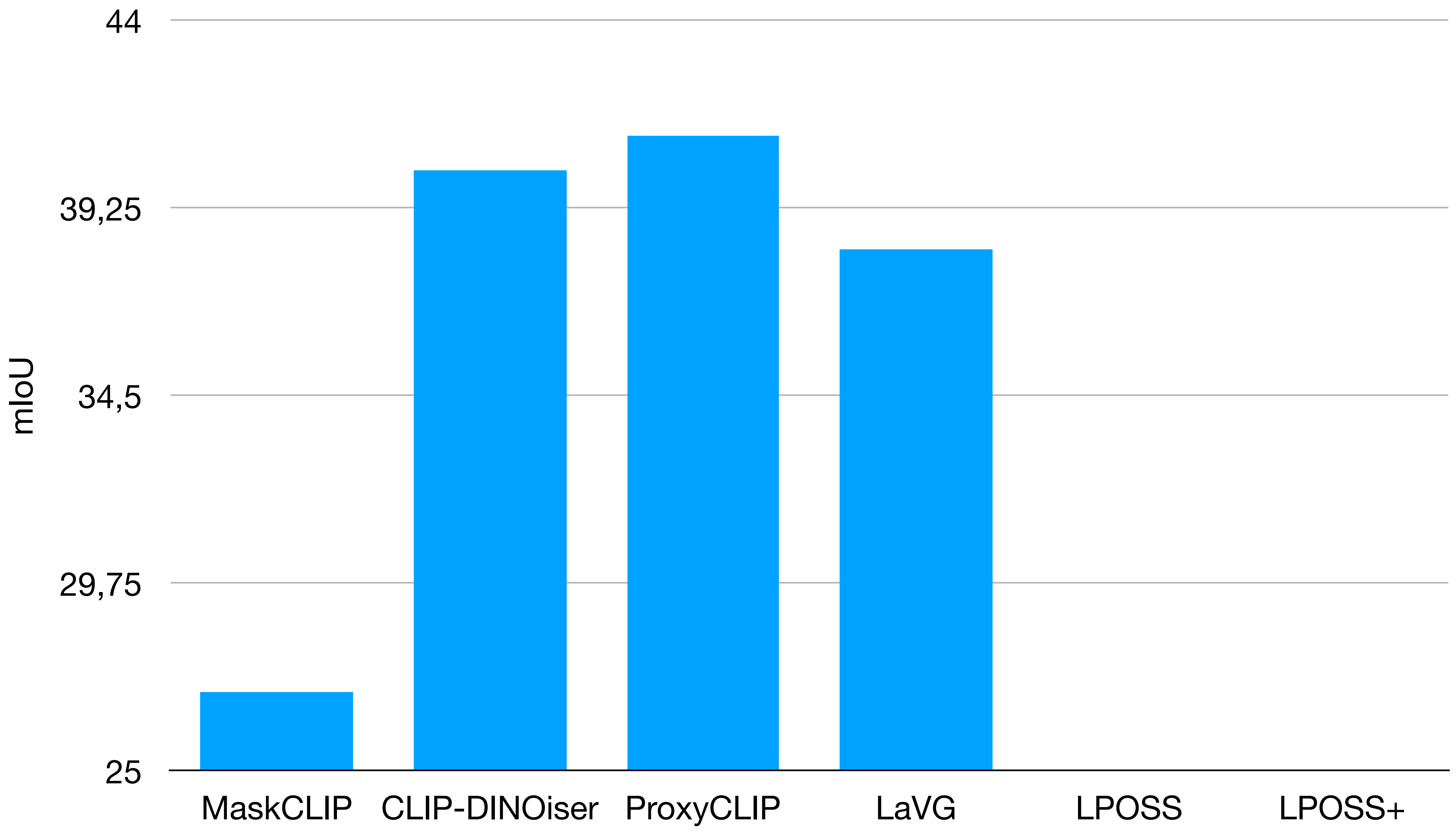
LPOSS

$Y^{(w_1)}_{\text{vlm}}$
VLM
sea
sand
class names
VLM
$Y^{(w_2)}_{\text{vlm}}$
upsample
$Y^{(img)}_{\text{lposs}}$
combine
upsample
prediction
argmax
segmentation

LPOSS

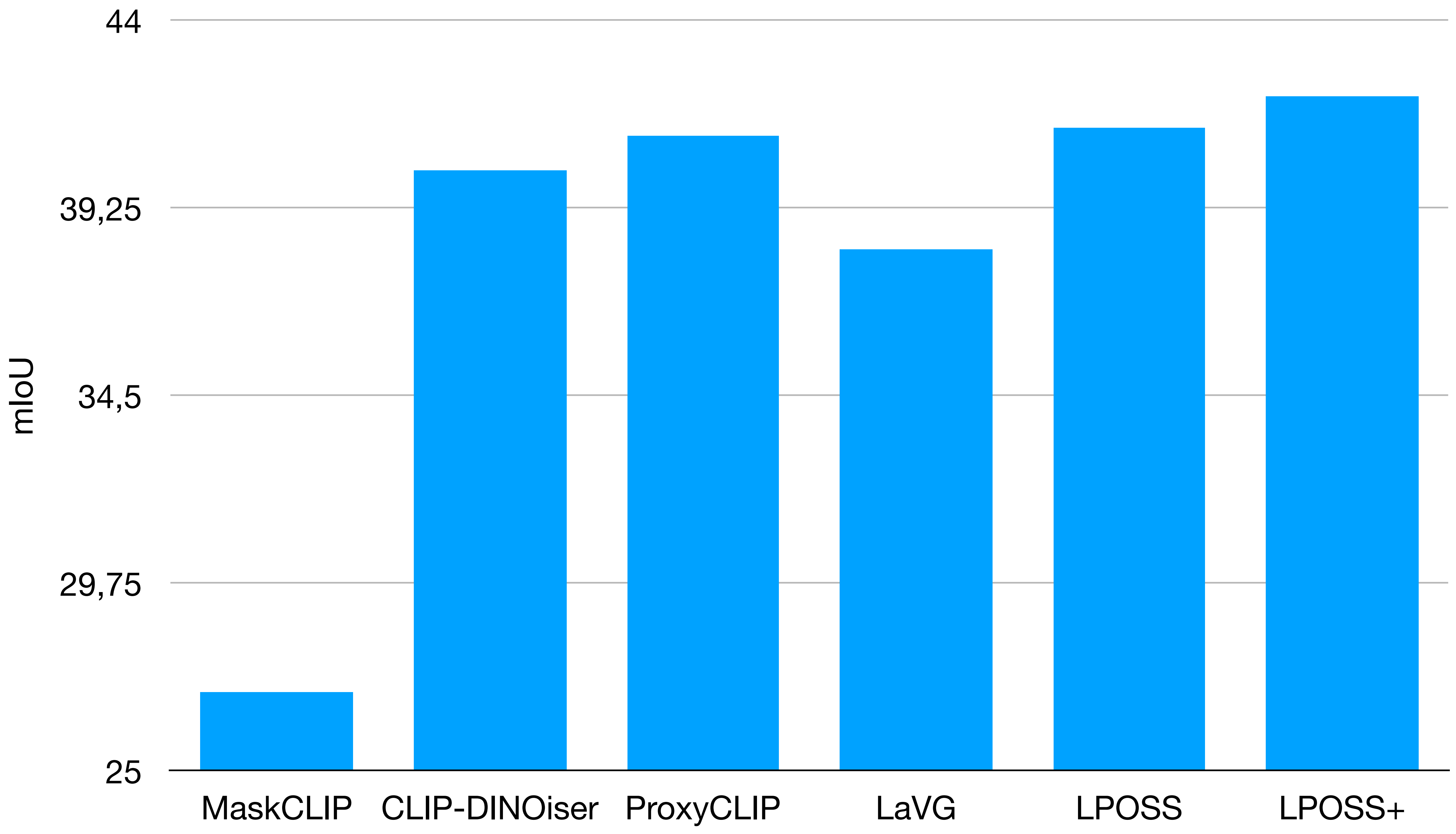
VM
VLM
VLM
VM
$Z^{(w_1)}_{\text{vm}}$
$Y^{(w_1)}_{\text{vlm}}$
$Y^{(w_2)}_{\text{vlm}}$
$Z^{(w_2)}_{\text{vm}}$
sea
sand
class names
LPOSS
$Y^{(w_1)}_{\text{lposs}}$
$Y^{(w_2)}_{\text{lposs}}$
upsample
upsample
combine
$Y^{(img)}_{\text{lposs}}$
prediction
argmax
segmentation

Results



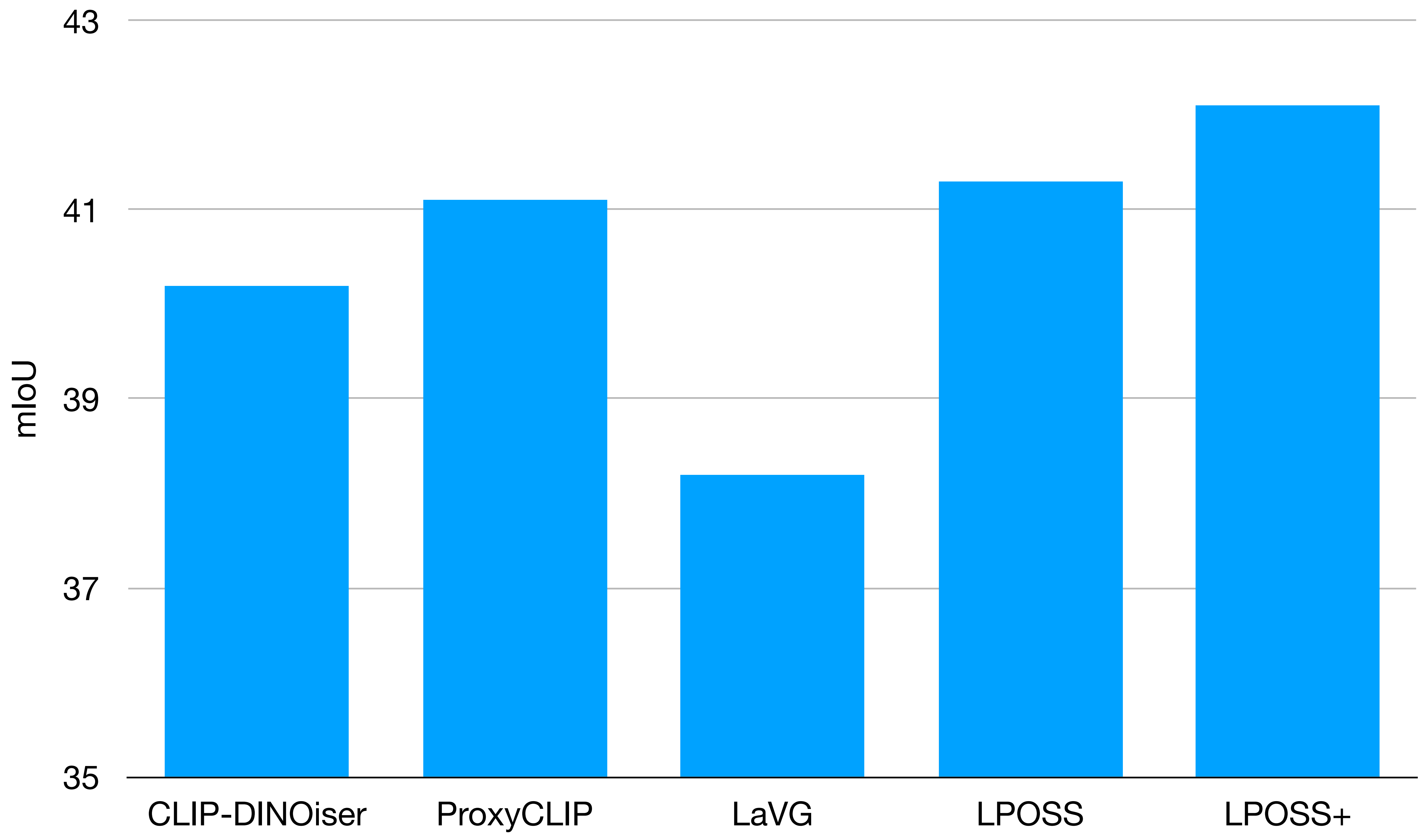
Averaged over 8 datasets

Results



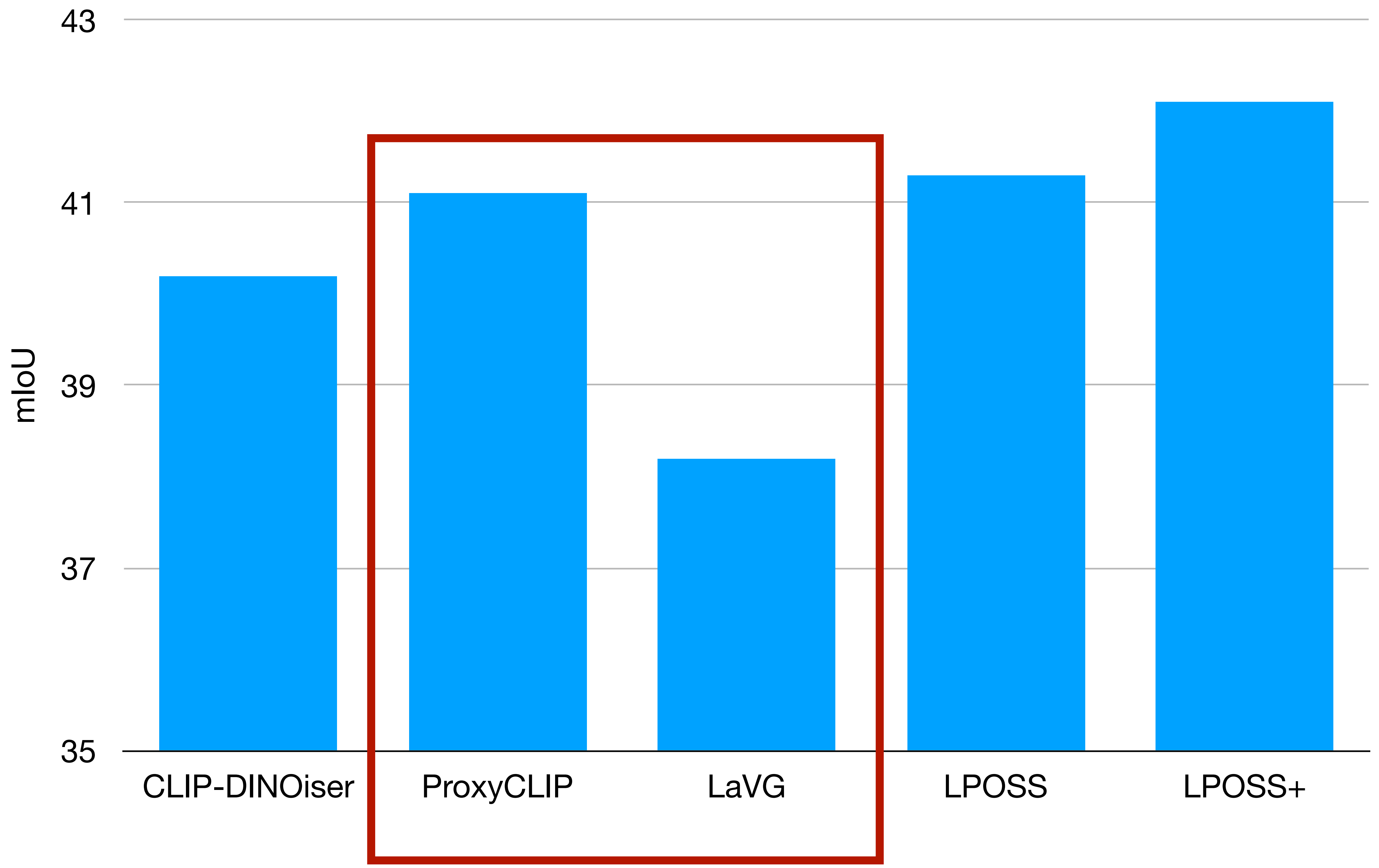
Averaged over 8 datasets

Results



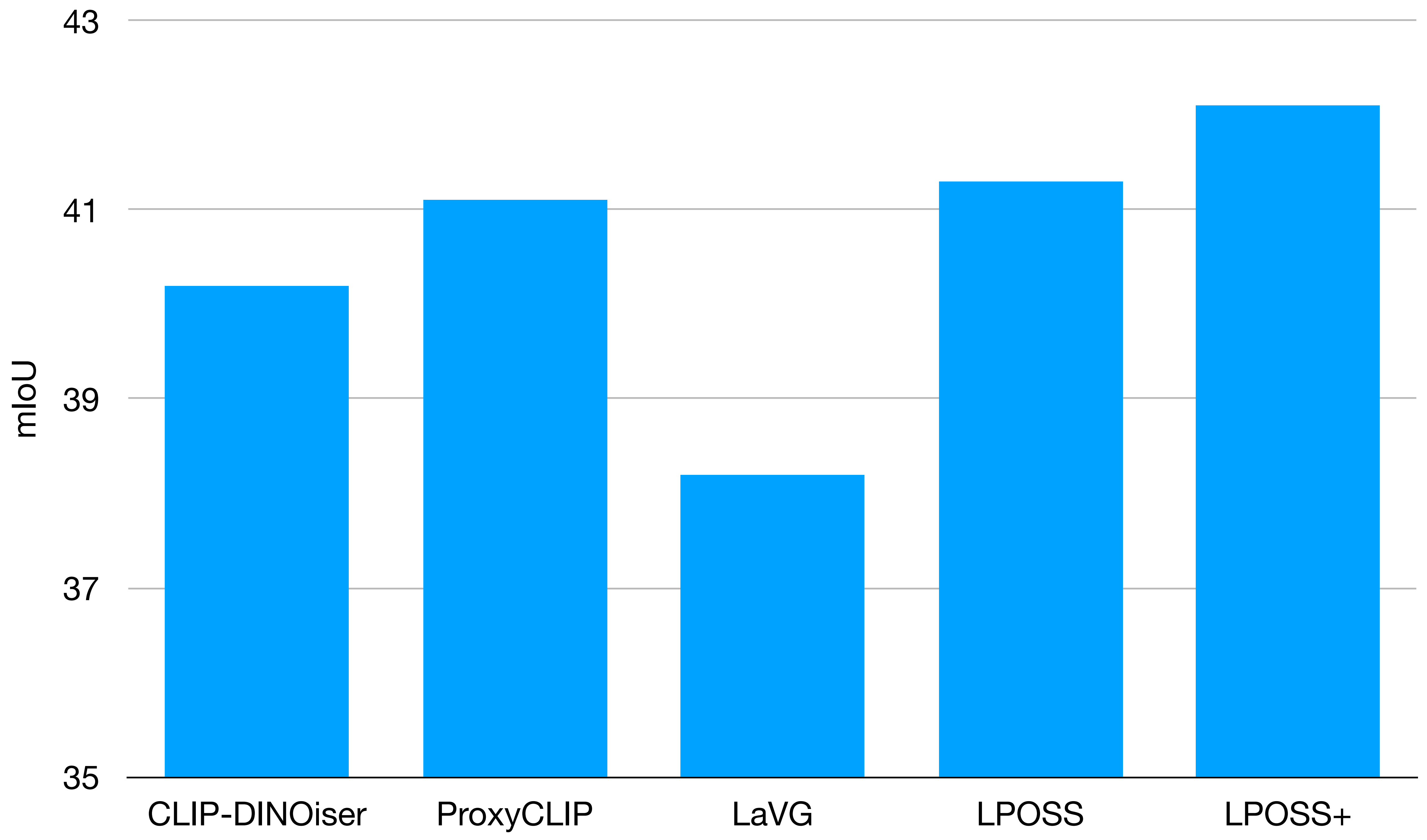
Averaged over 8 datasets

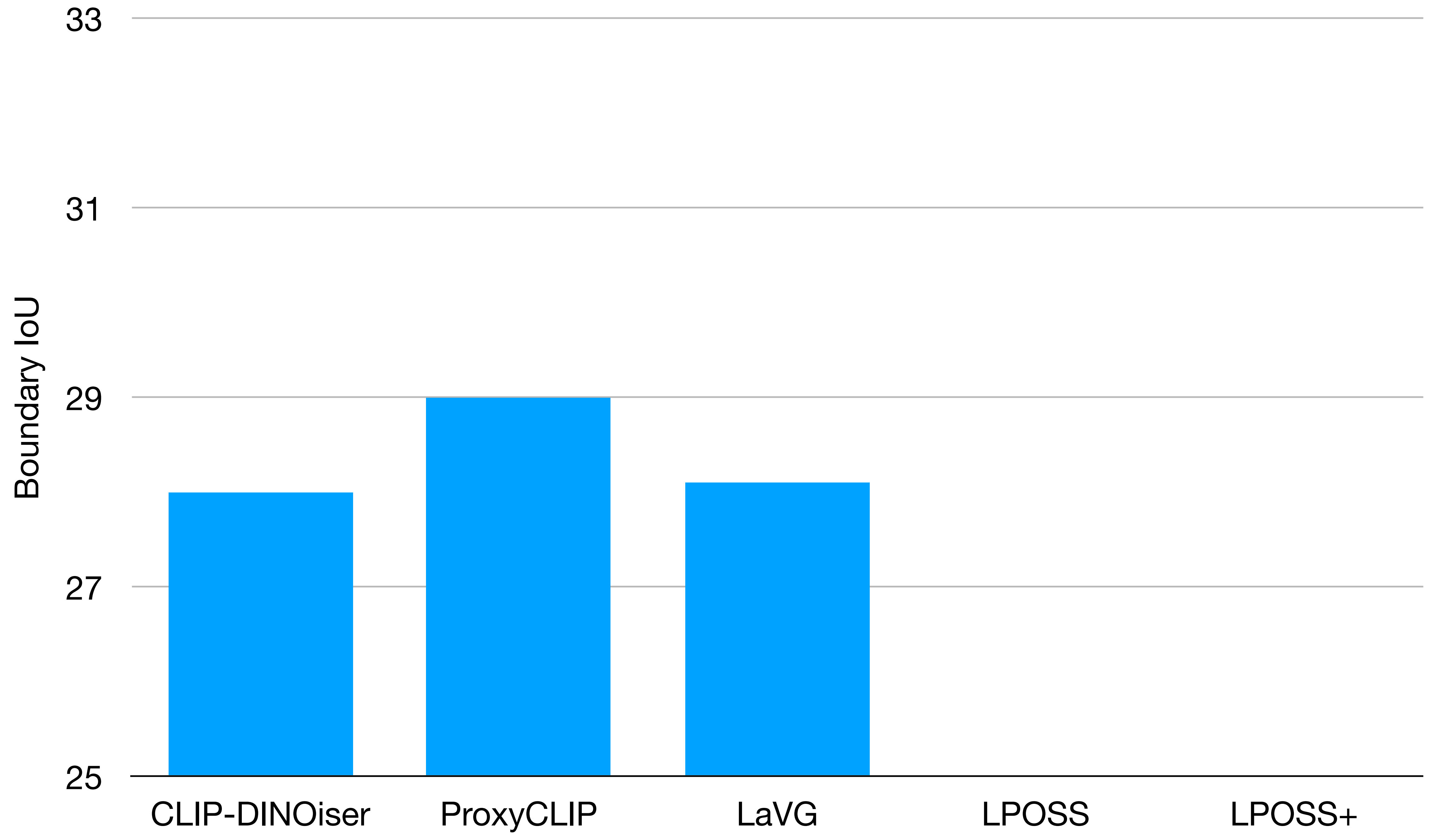
Results



Averaged over 8 datasets

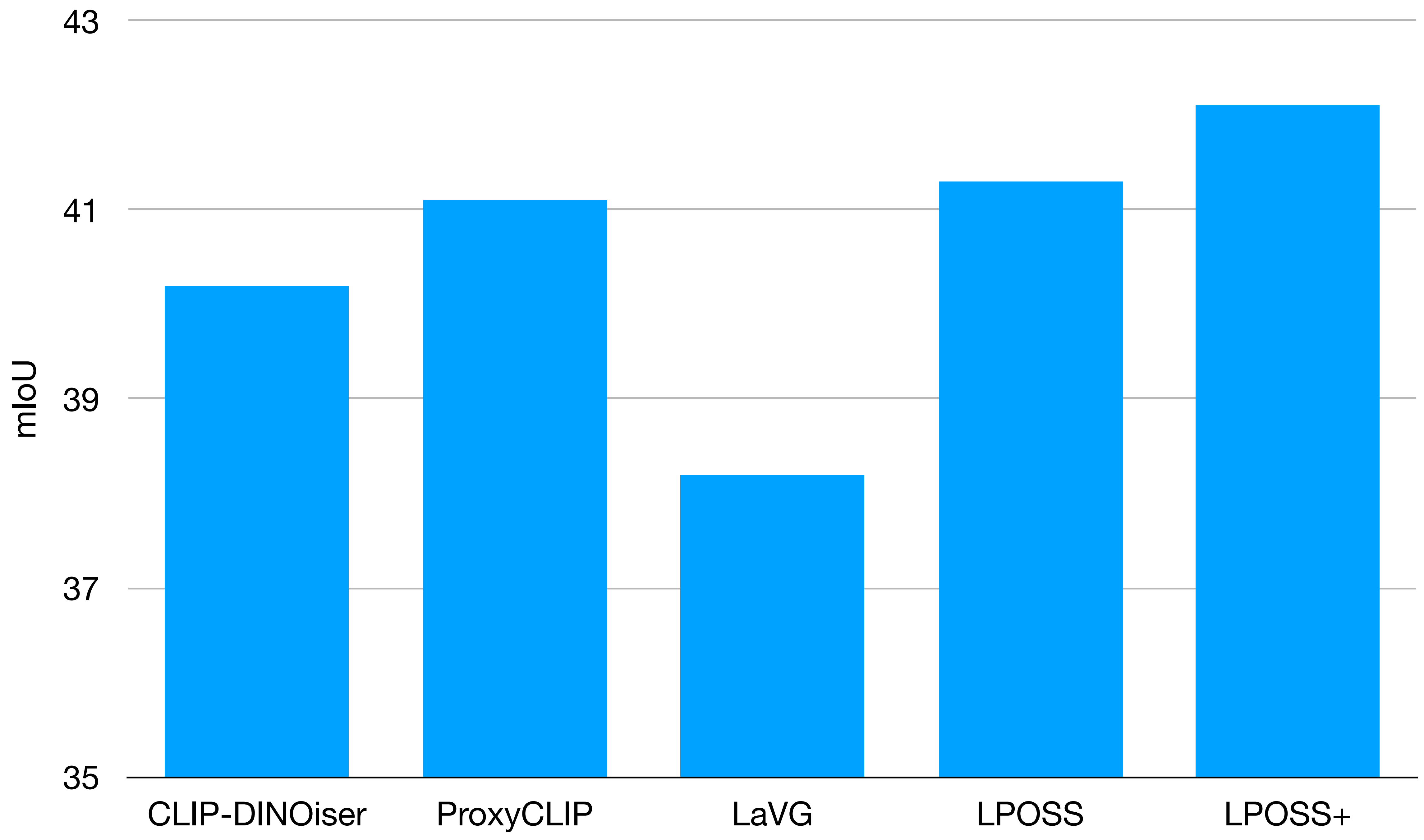
Results

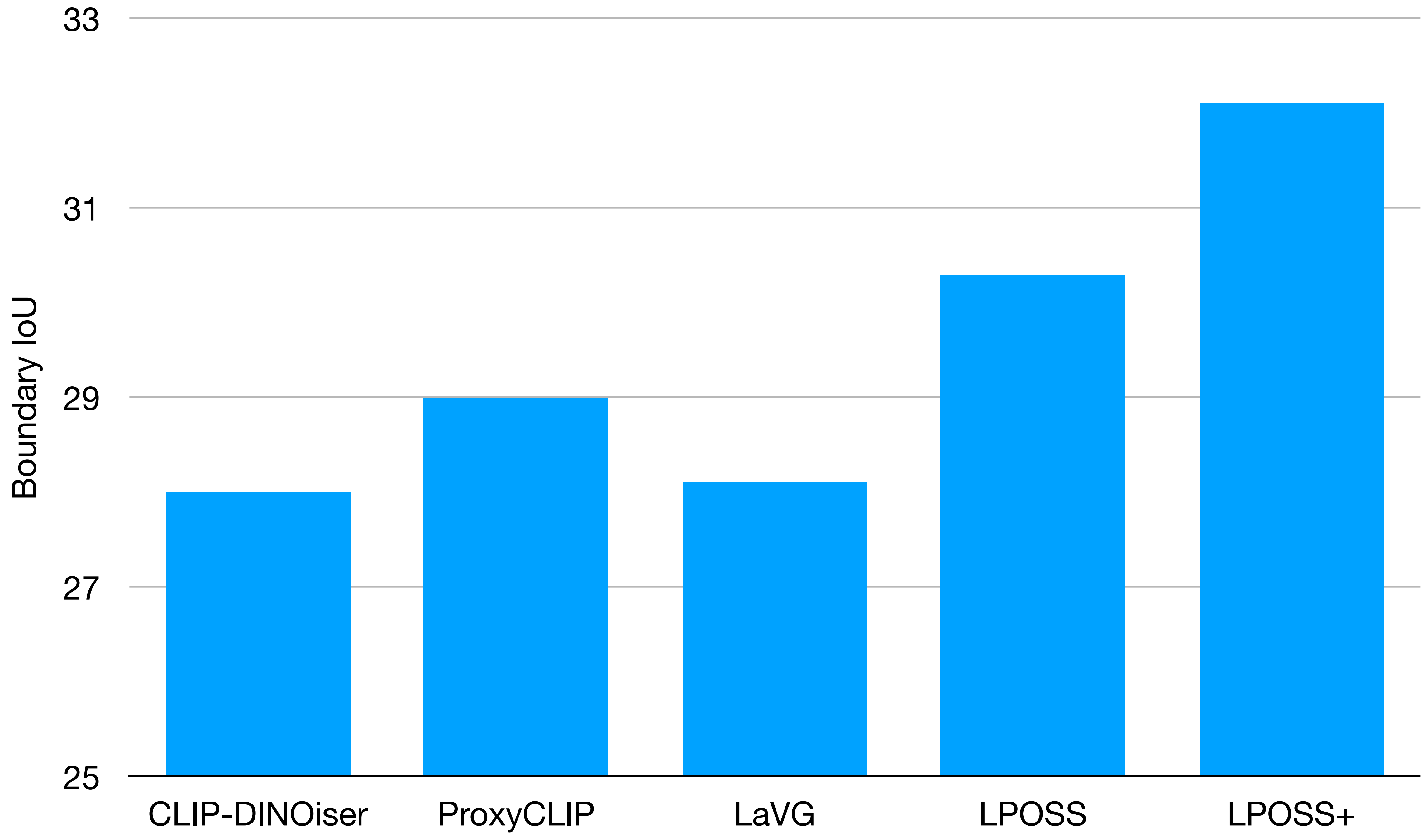




Averaged over 8 datasets

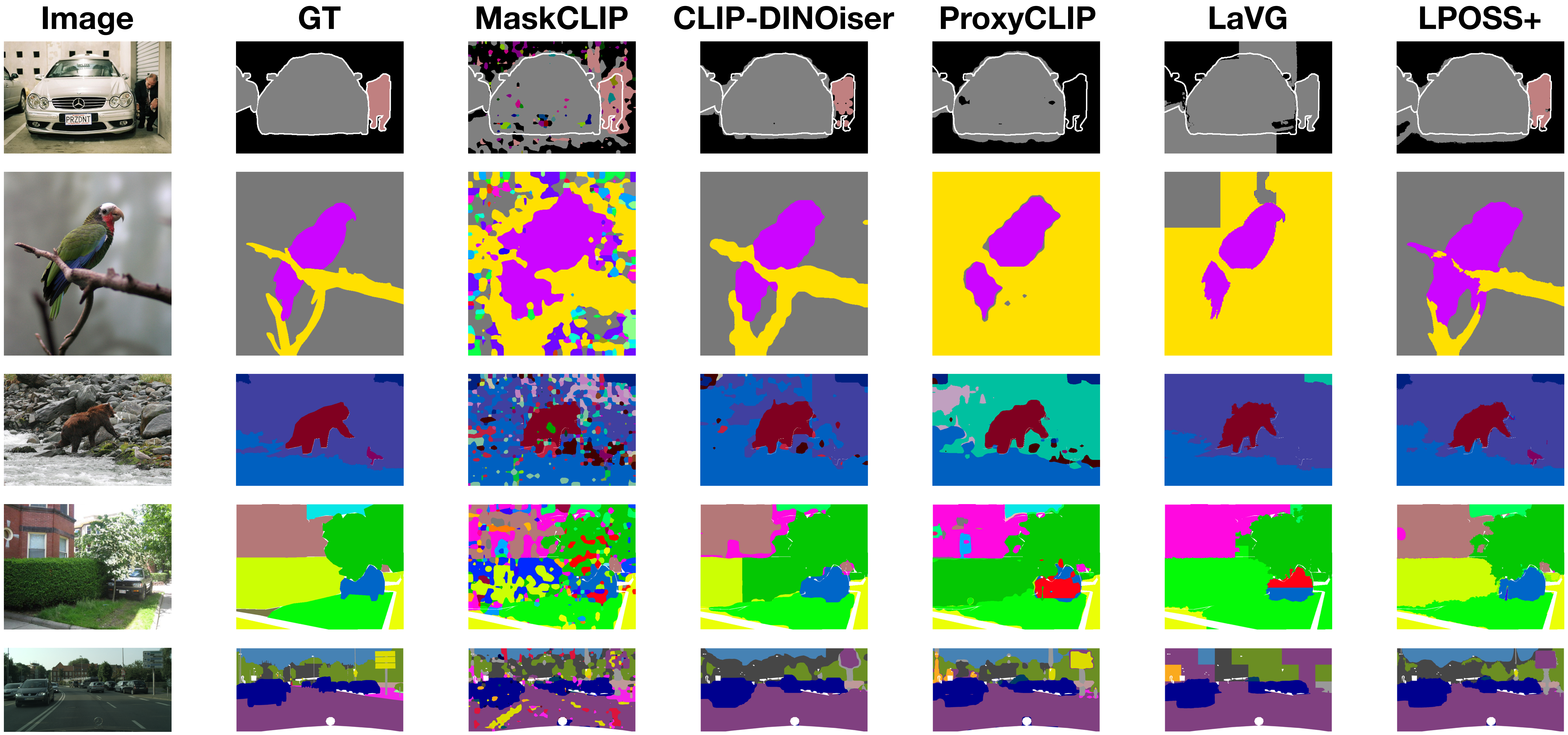
Results





Averaged over 8 datasets

Results



Approaches

- Training free methods

Approaches

- Training free methods
  - Hand designed on top of VLMs
  - MaskCLIP, LPOSS, etc.

Approaches

- Training free methods
  - Hand designed on top of VLMs
  - MaskCLIP, LPOSS, etc.
- Training on pixel-level annotations, but keep open-vocabulary ability

[1] Seokju Cho, Heeseong Shin, Sunghwan Hong, et.al. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. In CVPR, 2024.

[2] Bin Xie, Jiale Cao, Jin Xie, et.al. SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation. In CVPR, 2024.

Approaches

- Training free methods
  - Hand designed on top of VLMs
  - MaskCLIP, LPOSS, etc.
- Training on pixel-level annotations, but keep open-vocabulary ability
  - Fine-tune VLMs and train additional blocks on top
  - CAT-Seg [1], SED [2], etc.

[1] Seokju Cho, Heeseong Shin, Sunghwan Hong, et.al. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. In CVPR, 2024.

[2] Bin Xie, Jiale Cao, Jin Xie, et.al. SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation. In CVPR, 2024.

Evaluation

- Training on COCO (Stuff, Panoptic, …)
- Standard test sets
  - PASCAL (VOC and Context)
  - ADE20k
  - Cityscapes

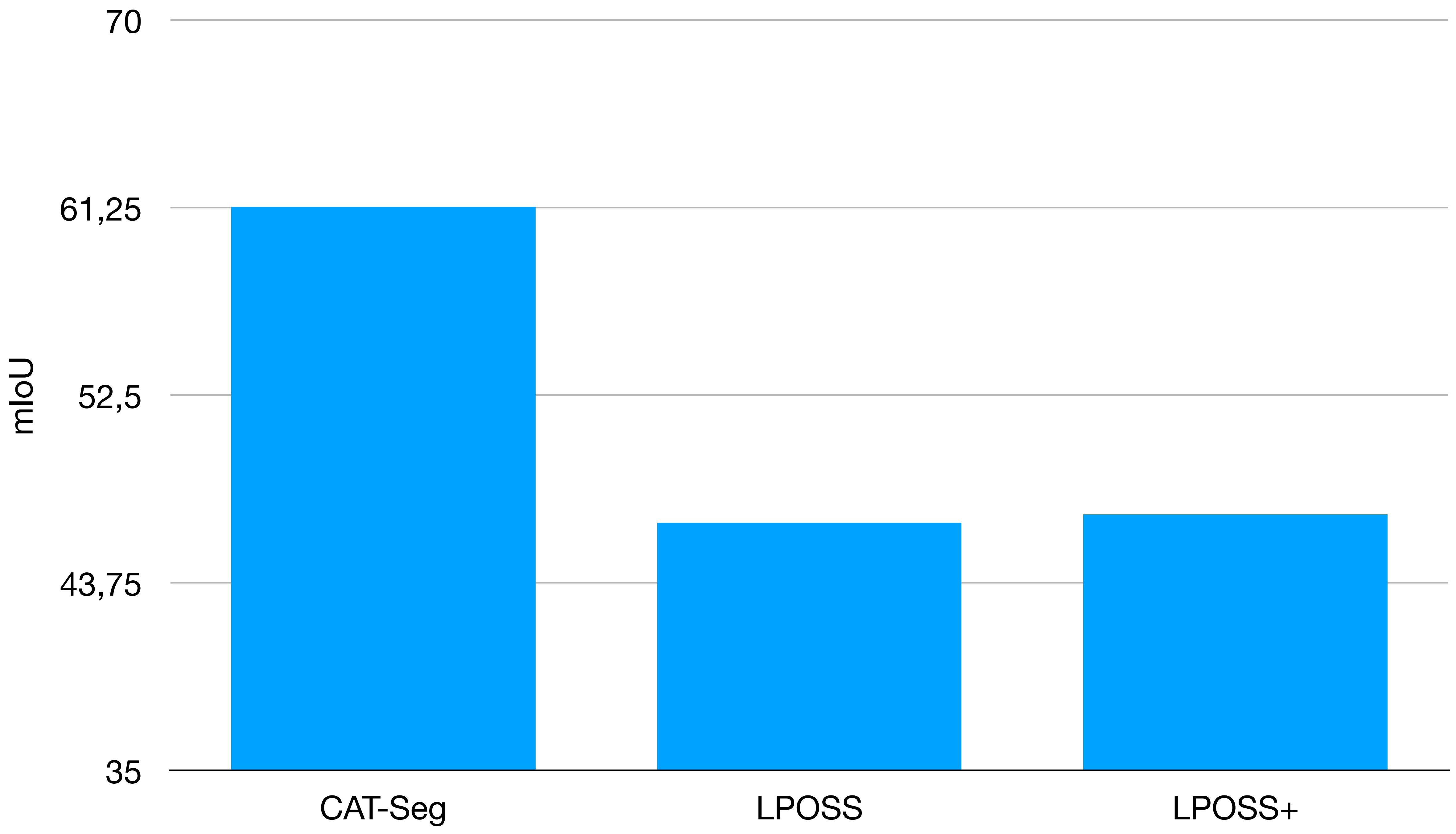
Evaluation

- Training on COCO (Stuff, Panoptic, …)
- Standard test sets
  - PASCAL (VOC and Context)
  - ADE20k
  - Cityscapes
- Potentially a large overlap with classes used in training

MESS benchmark [1]

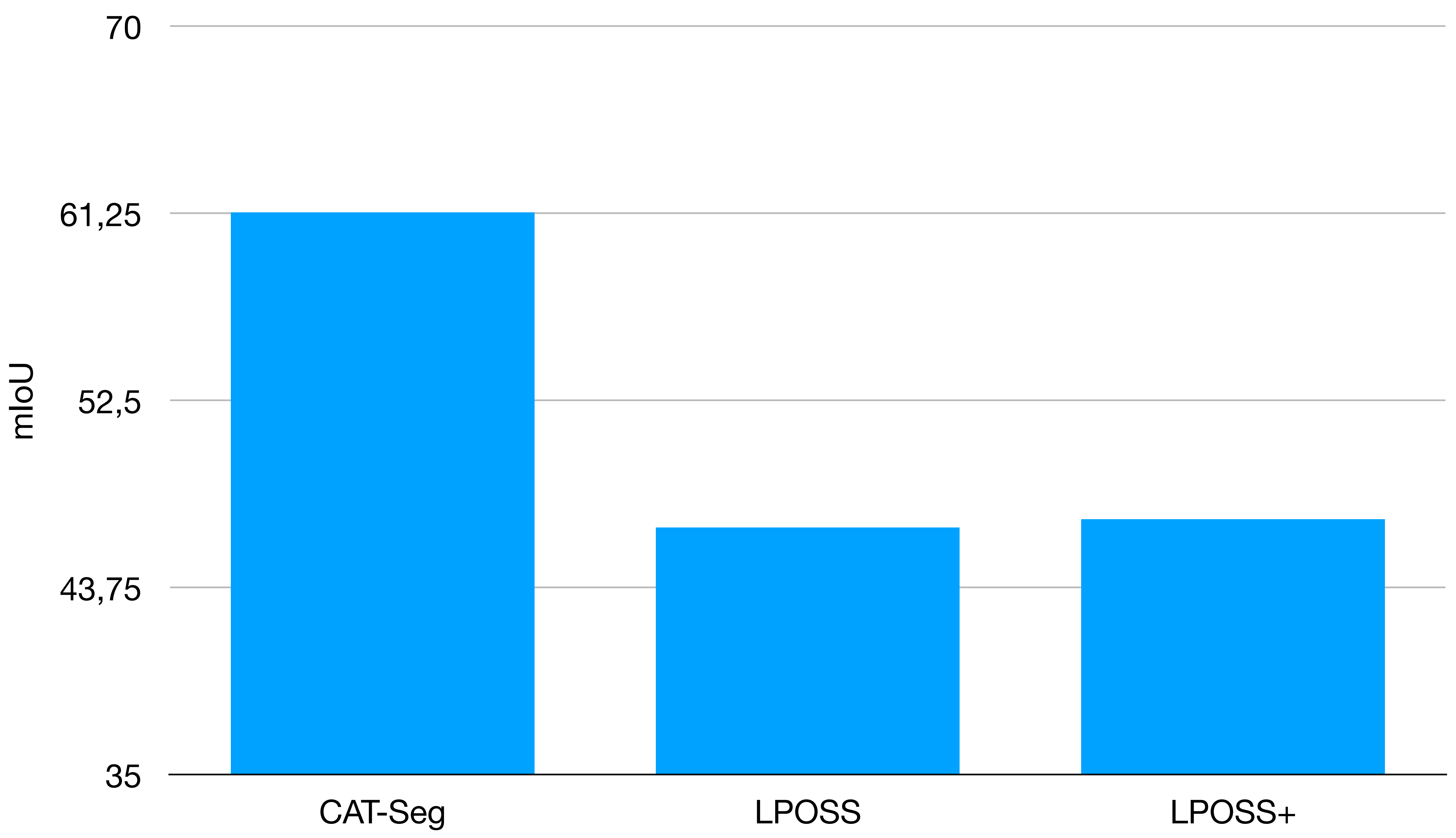
[1] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kuhne, Michael Vossing. What a MESS: Multi-Domain Evaluation of Zero-Shot Semantic Segmentation. In NeurIPS, 2023.

Results

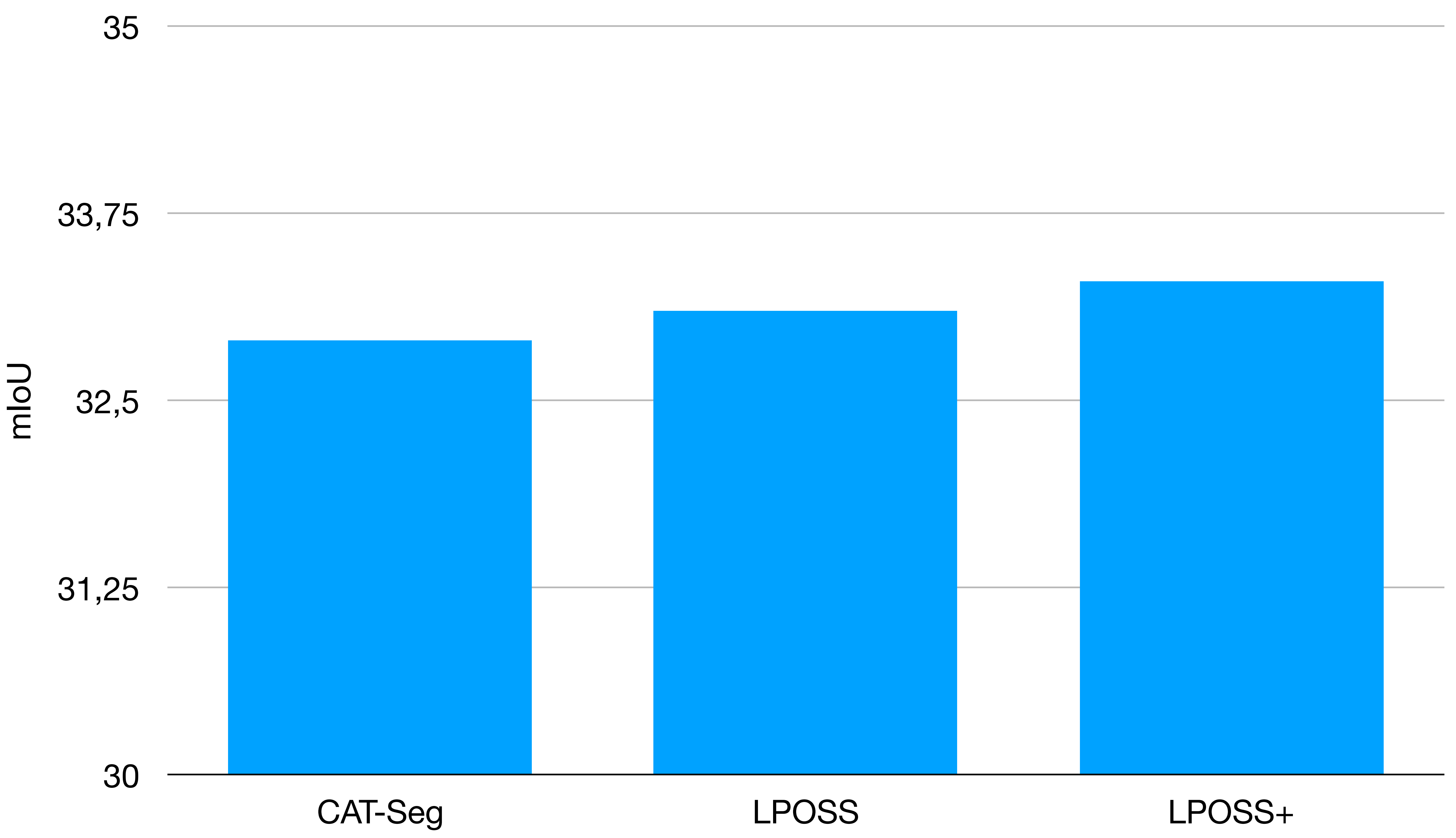


Averaged over 3 standard datasets
Close to the training set distribution

Results



Averaged over 3 standard datasets
Close to the training set distribution



Averaged over 22 MESS datasets
Very diverse test sets

Demo