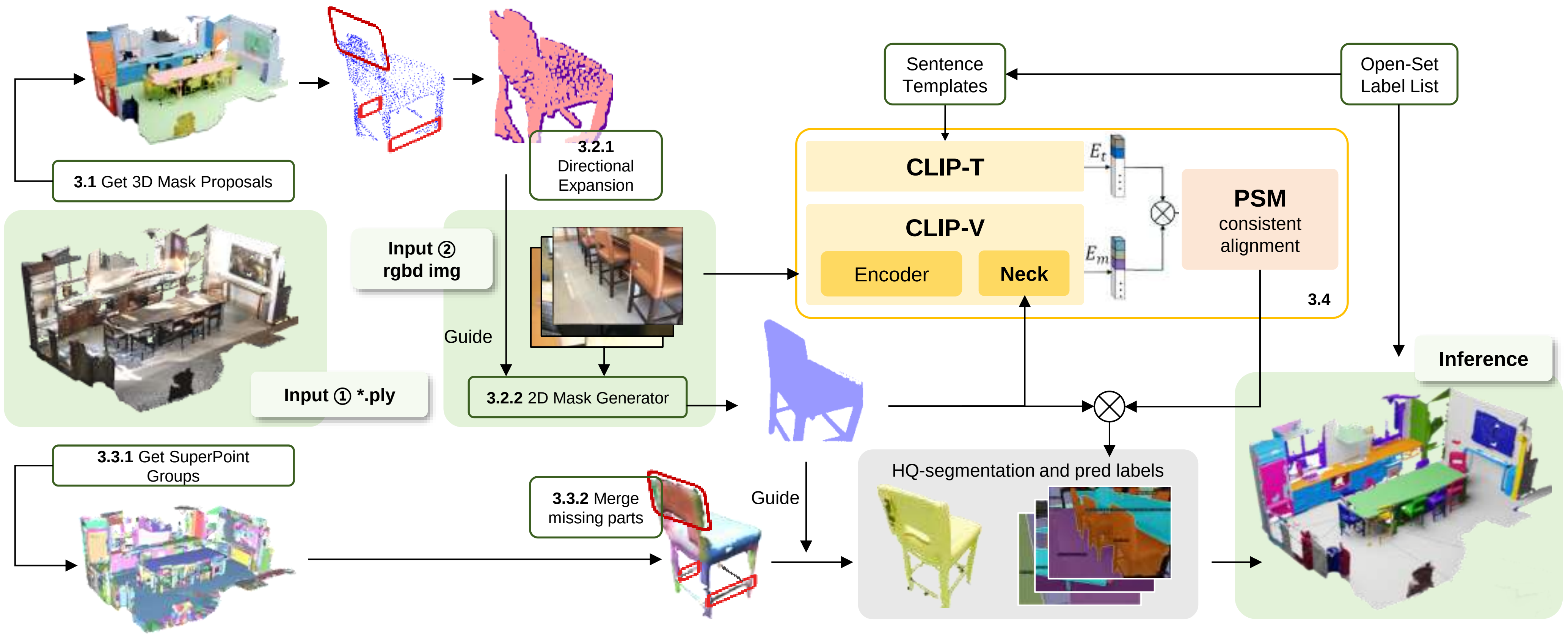
MOSS: Mask-Oriented Open-Set for 3D Scene Segmentation using Superpoint

Nov 2024 - ?



Task Background

- Traditional methods rely on densely annotated 3D scenes.
- Have to utilize supervision from ground truth labels.
- 3D Annotation is **Time-consuming** and **Expensive!**

Input 3D
Geometry

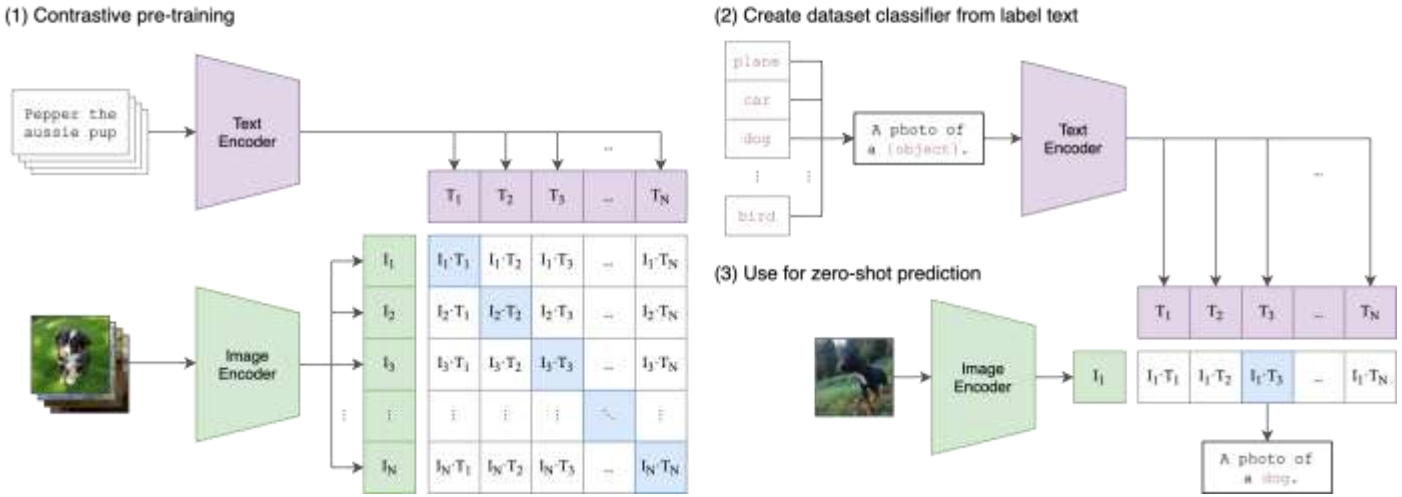
Annotated
3D scenes

Task Background

Meanwhile computer vision is going through a transition from the previous **closed-set** perception to **open-set** perception:

Closed-set: only handles predefined classes during training and has limited capability in dynamic world

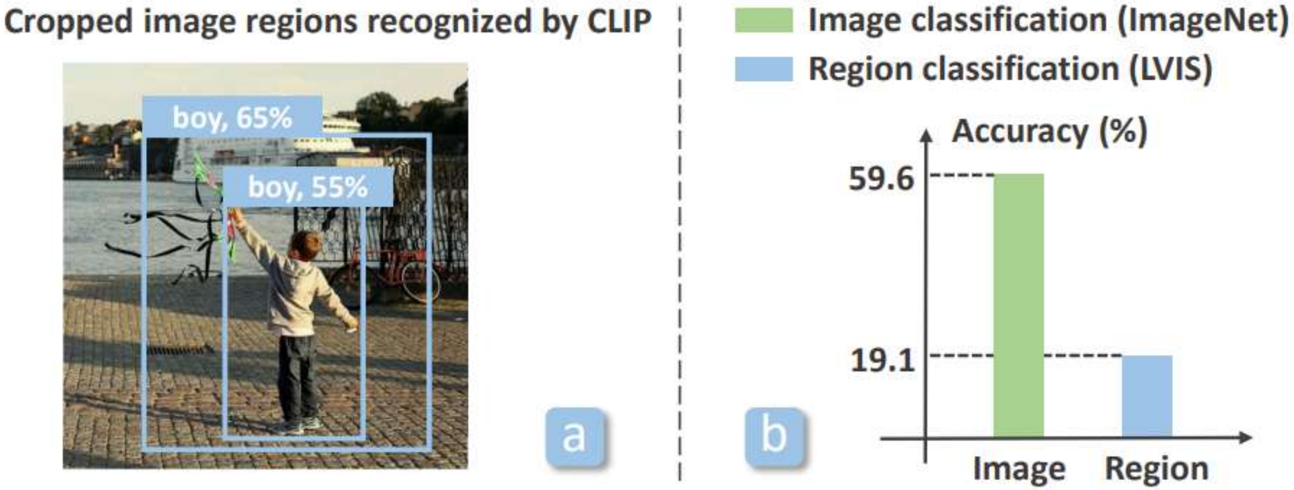
Open-set: understands unseen, diverse and free-flowing language, mimicking how humans naturally interact with the world and each other

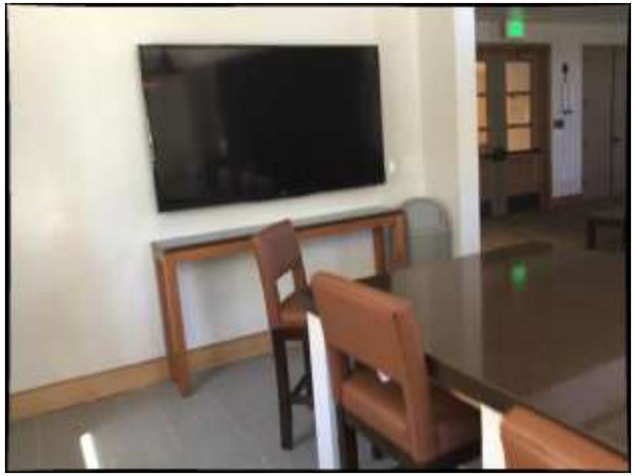


2D open-set tasks can now understand new concepts, perform accurate segmentation and detection, and handle complex tasks requiring reasoning.

Motivation

- Previous methods simply use the original CLIP model without addressing its regional limitations.
- Directly applying it for object detection leads to poor performance due to domain shift, as CLIP was trained to match whole images to text descriptions, without capturing fine-grained alignment between image regions and text spans.



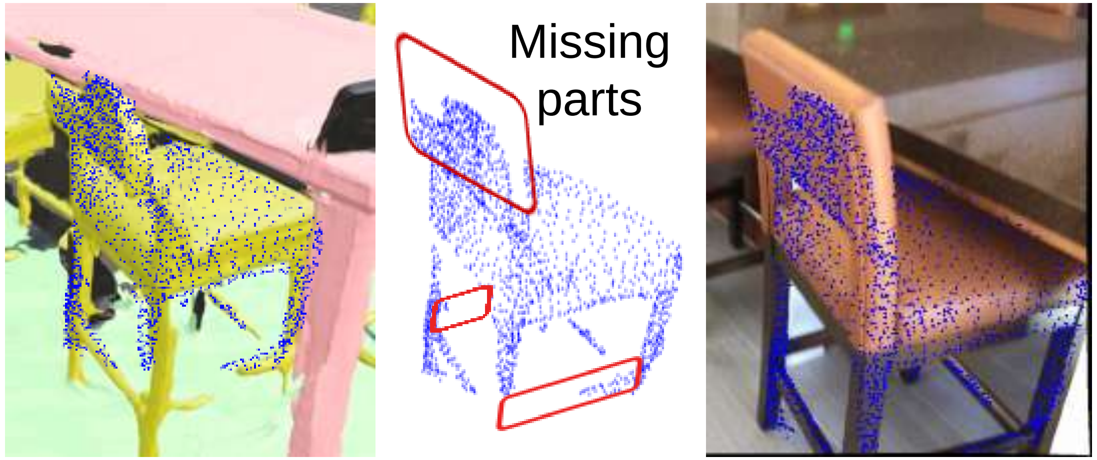


Y. Zhong et al., RegionCLIP: Region-based language-image pretraining, CVPR,2022.

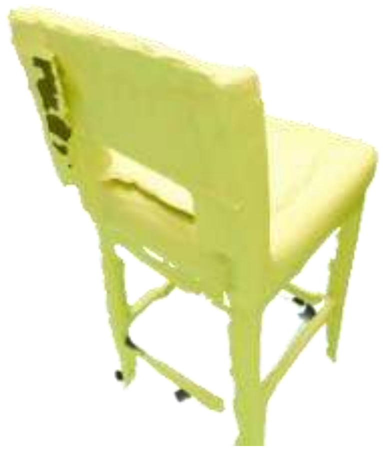
Motivation

- Most existing work heavily relies on the mask proposals generated by pre-trained 3D models(like Mask3D), where the quality of these masks directly affects the performance of instance segmentation.
- However, open-set tasks should not be constrained by closed-set models. Additionally, prior knowledge from 2D segmentation models can alleviate the limitations observed in current 3D class performance.

Mask3D



3D Points to 2D Pixels

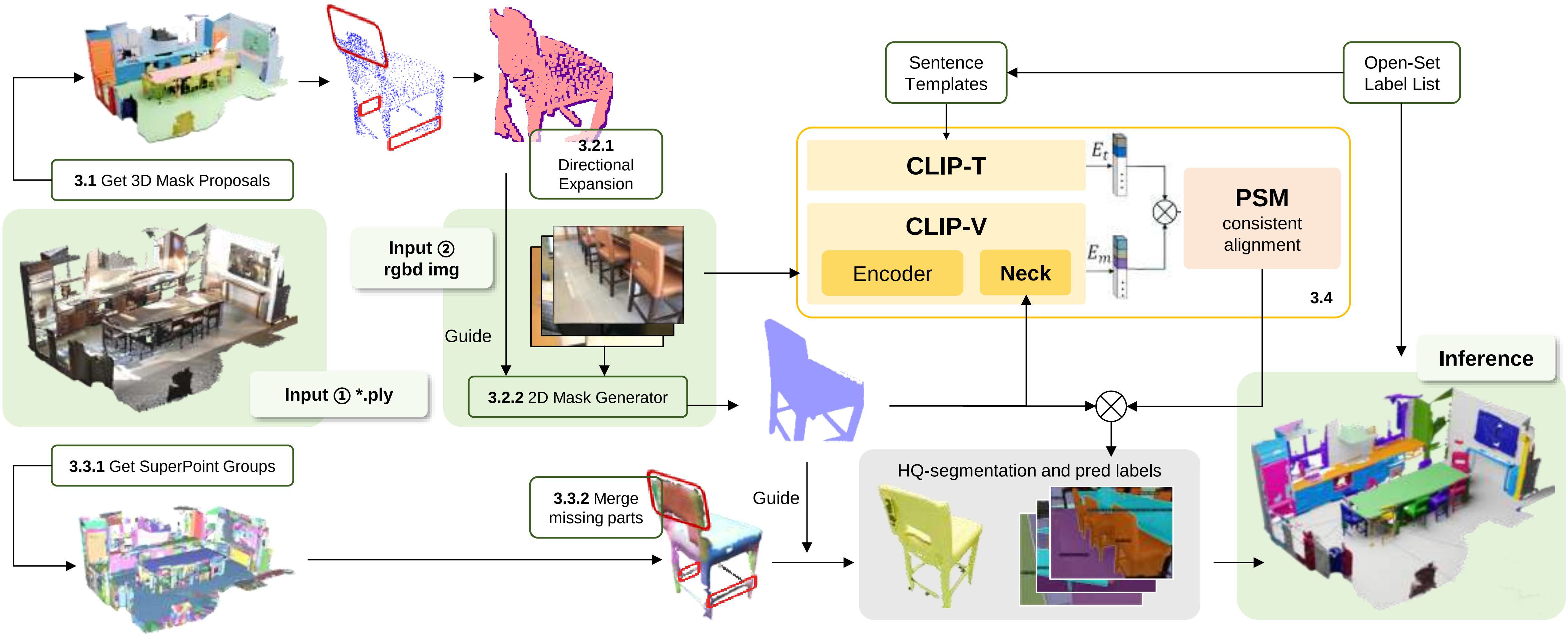


Ground truth

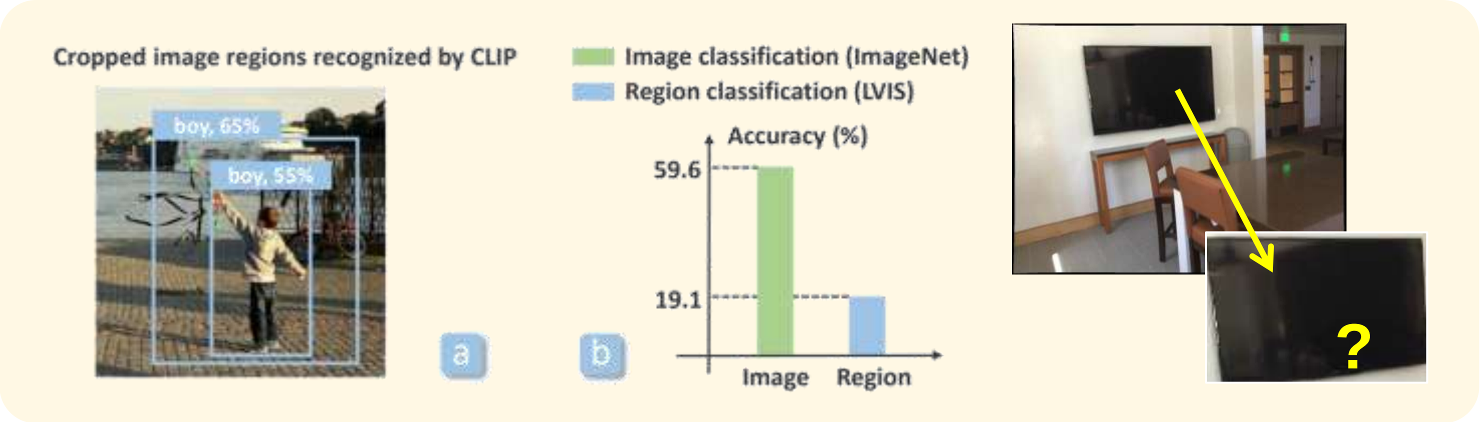
Contributions

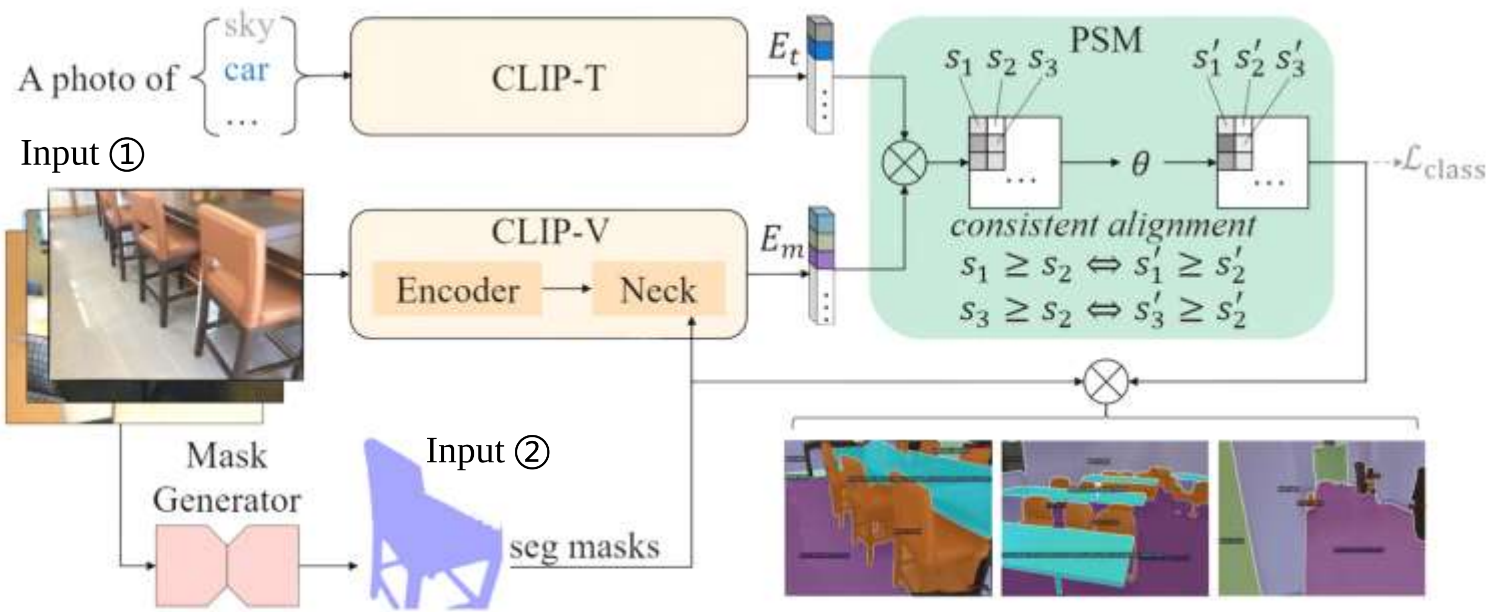
- We proposed **MOSS**, a mask-based framework for open-set 3D scene semantic segmentation that enables efficient cross-dimensional feature transfer and inference.
- We enhanced the frame by implementing **global information input with mask** constraints to strengthen attention.
- We employed **a density-guided dilation algorithm** to optimize the matching precision between 2D and 3D masks.
- We also introduce a novel method to enhance 3D mask proposals, which leverages 2D prior knowledge to perform **back-projection** on a 3D pre-trained model. This approach guides **the capture of superpoint clusters** in the 3D scene, thereby improving the quality of the output results of **fine-tuning the close-set model result**.

Our Proposed MOSS



Our Proposed MOSS - Contribution1





Input a cropped image:

- Cropped regions lack global information.

Input a purely global image:

- Fail to localize the areas that need to be understood.
- Lead to inconsistencies in the granularity of classification, where the level of detail may be too fine or too coarse to align with the designated regions of the mask proposal.

Input both global images and masks:

- Constrain regions requiring enhanced understanding.
- Obtain contextual information to improve inference accuracy.

Zeng et al. MaskCLIP++, a mask-based CLIP fine-tuning framework for open-vocabulary image segmentation, 2024

- **How to obtain a high-quality 2D mask?**

Our Proposed MOSS - Contribution2

Density-based Directional Expansion Algorithm

IOU?

The number of pixels does not correspond

① Expansion in the equal direction

② Directional expansion based on density

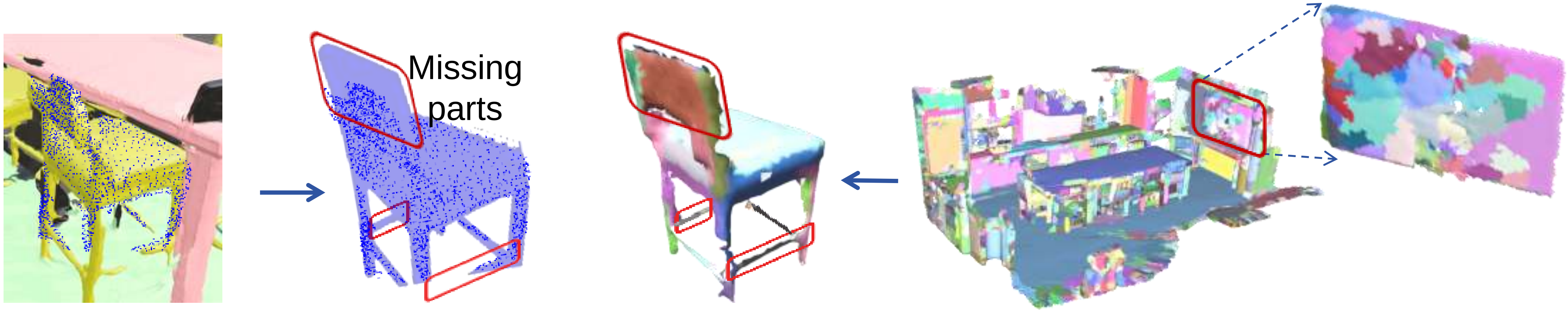
③ ...

④ Calculate IoU > the threshold

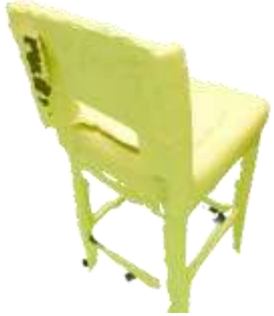
- **What else can a high-quality 2D mask offer?**

Our Proposed MOSS - Contribution3

Fine-Tuning of 3D Mask Proposals based on SuperPoints



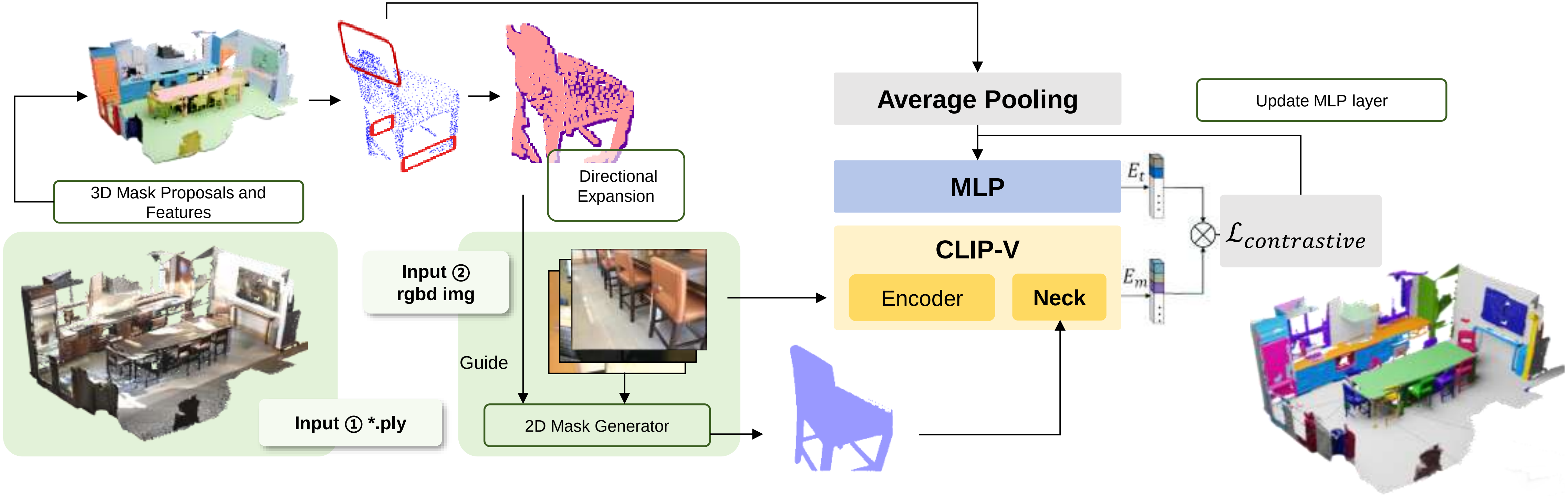
- The coarse 3D mask proposal has missing areas compared to the ground truth.
- The prior knowledge from the 2D can fill these gaps.
- Pixels from the 2D mask are projected to 3D, and the matching points are added to the 3D mask.


Ground truth

- To improve efficiency, the raw point cloud is transformed into superpoint clusters.
- SuperPoints: points are grouped into geometrically homogeneous regions.
- Instead of individual points, superpoint clusters are used as the unit for merging.

- **Can we accelerate the Inference Process?**

Our Proposed MOSS – Contribution4



Knowledge distillation:

- We first obtain the 3D point cloud features from the output of Mask3D
- For each 3D proposal, we apply average pooling to derive its 3D feature vector
- We then use a multilayer perceptron (MLP) to map it into the image-text embedding space of CLIP
- To train the MLP layer, we employ a contrastive loss function on the dataset, optimizing the alignment between the 3D and 2D mask embeddings.

Experiments & Results (Imcomplete)

Method	mAP	mAP50	mAP25	head	comm	tail
Mask3D (Closed Vocab.)	26.9	36.2	41.4	39.8	21.7	17.9
SAM3D	6.1	14.2	21.3	7	6.2	4.6
OVIR-3D	13	24.9	32.3	14.4	12.7	11.7
Open3DIS	23.7	29.4	32.8	27.8	21.2	21.8
OpenScene (2D Fusion)	11.7	15.2	17.8	13.4	11.6	9.9
OpenScene (3D Distill)	4.8	6.2	7.2	10.6	2.6	0.7
OpenScene (2D-3D Ens.)	5.3	6.7	8.1	11	3.2	1.1
OpenMask3D	15.4	19.9	23.1	17.1	14.1	14.9
OpenMask3D	16.2	21.3	24.8	22.2	13.4	12.5
Open3DIS	18.6	23.1	27.3	24.7	16.9	13.3
Open-YOLO 3D	<u>24.7</u>	<u>31.7</u>	<u>36.2</u>	<u>27.8</u>	<u>24.3</u>	<u>21.6</u>
MOSS(Ours)	**27.1**	**36.3**	**41.7**	**30.6**	**25.9**	**24.6**

* The experimental results are still being updated
(hyperparameters are being finalized)

Phase Summary & Next Work

- Tasks in the 2D domain can successfully guide 3D spatial understanding tasks;
- The instance matching between 2D and 3D in the form of masks works effectively

- Conduct experiments with additional datasets;
- Attempt to replace the black-box CLIP model with an integrated VLM (Visual-Language Model) during the inference phase;
- Optimize code details to reduce the inference time per scene.