Cross-Modal and Uncertainty-Aware Agglomeration for Open-Vocabulary 3D Scene Understanding

Jinlong Li[1], Cristiano Saltori[1], Fabio Poiesi[2], Nicu Sebe[1]

[1]University of Trento [2]Fondazione Bruno Kessler

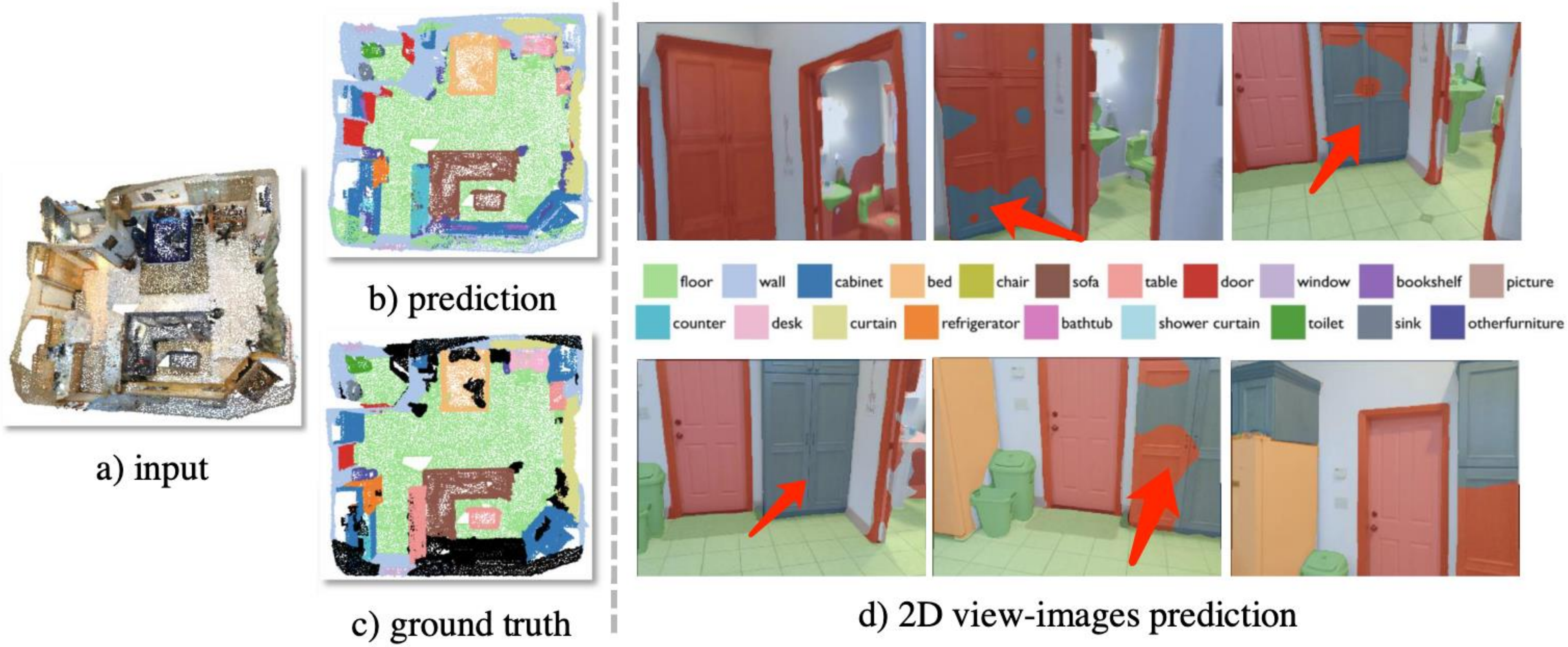
Outline

- Motivation
- Method
- Experiments

Motivation

Given the multi-view posed images available,
distill the 2D foundation model knowledge into 3D model by 2D-3D projection

But, noises cause feature embedding ambiguity, across consecutive image sequences



a) input b) prediction c) ground truth d) 2D view-images prediction

Preliminary study on image embedding ambiguity. VLM embeddings show inconsistent segmentations across multi-view images (e.g. cabinet). The guidance with ambiguous embeddings may be detrimental for supervising a 3D model training.

Motivation

Multiple 2D foundation models available, why not embrace them?

Model	Training Dataset	Dataset Size	Architecture	Objective
ViT [17]	ImageNet-1k/21k	1.2M/14.2M	ViT-B/L/G	Supervised classification
DINOv2 [61]	LVD-142M	142M	ViT-L/14	Discriminative self-supervised learning
CLIP [66]	WebImageText	400M	ViT-L/14	Image-text contrastive learning
Stable Diffusion [68]	LAION	5B	UNet	Image-Text/Image Generation

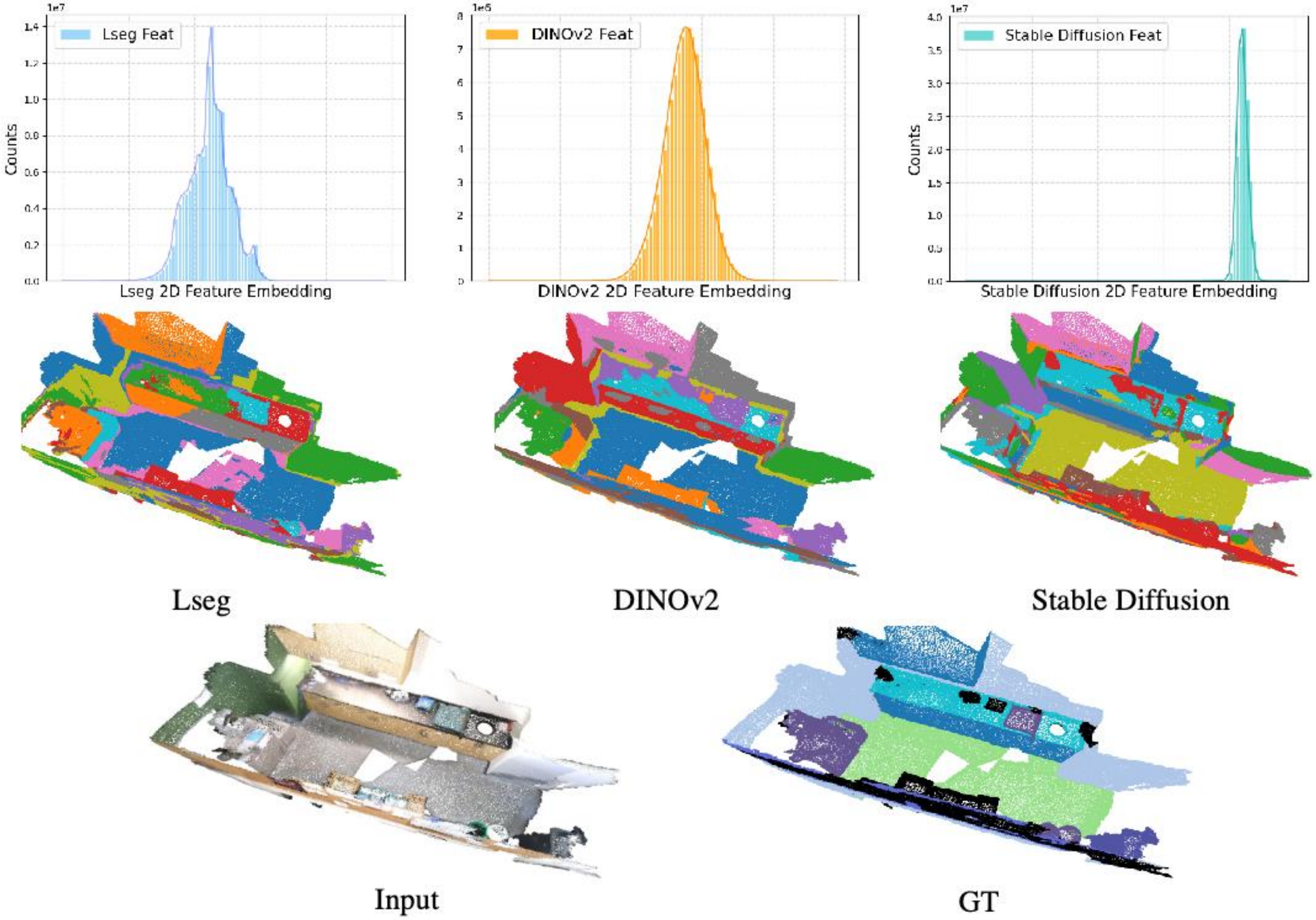
DINOv2 --- depth and surface normal

Diffusion Model --- geometric tasks

CLIP --- visual & textual multi-modal ability

[1] Probing the 3d awareness of visual foundation models
[2] Lexicon3d: Probing visual foundation models for complex 3d scene understanding

Motivation



Different 2D foundation models
heterogeneous & complementary

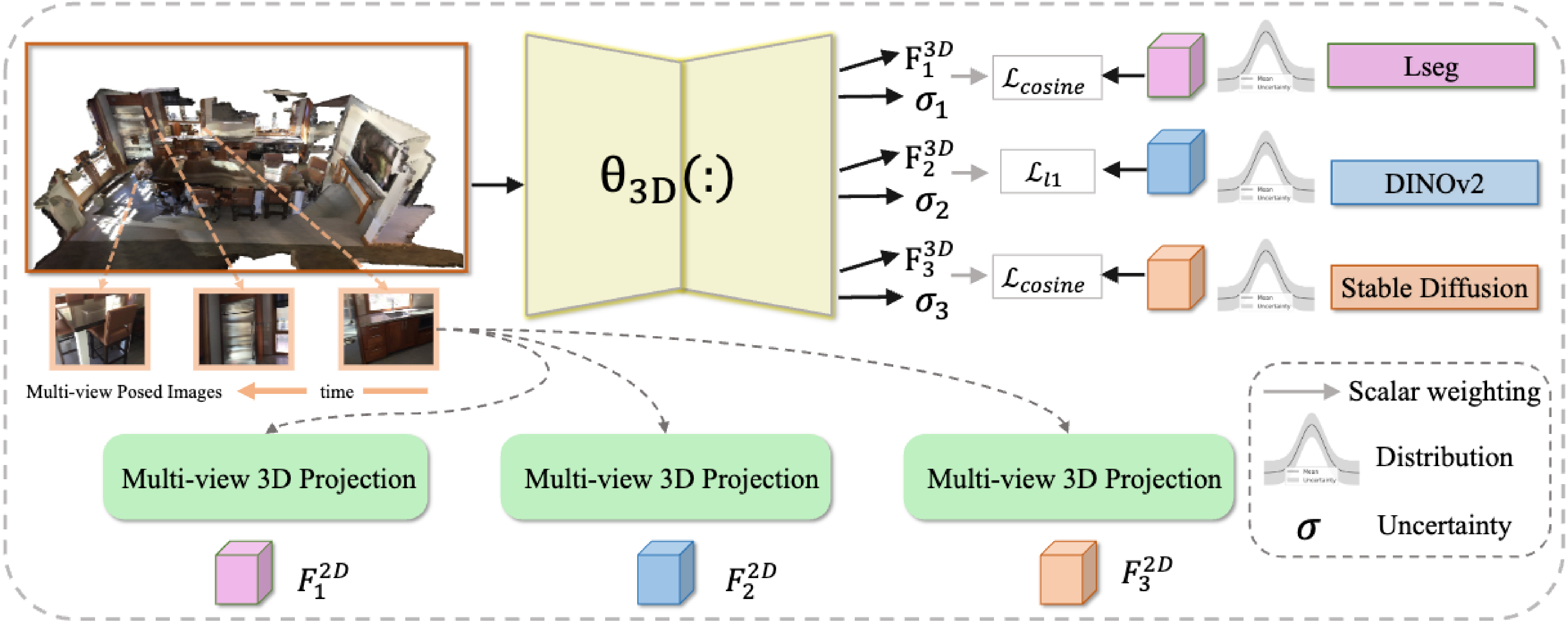
gaussian-like feature distribution

[3] Learning transferable visual models from natural language supervision
[5] Dinov2: Learning robust visual features without supervision
[4] High-resolution image synthesis with latent diffusion models
[6] Language-driven semantic segmentation

Outline

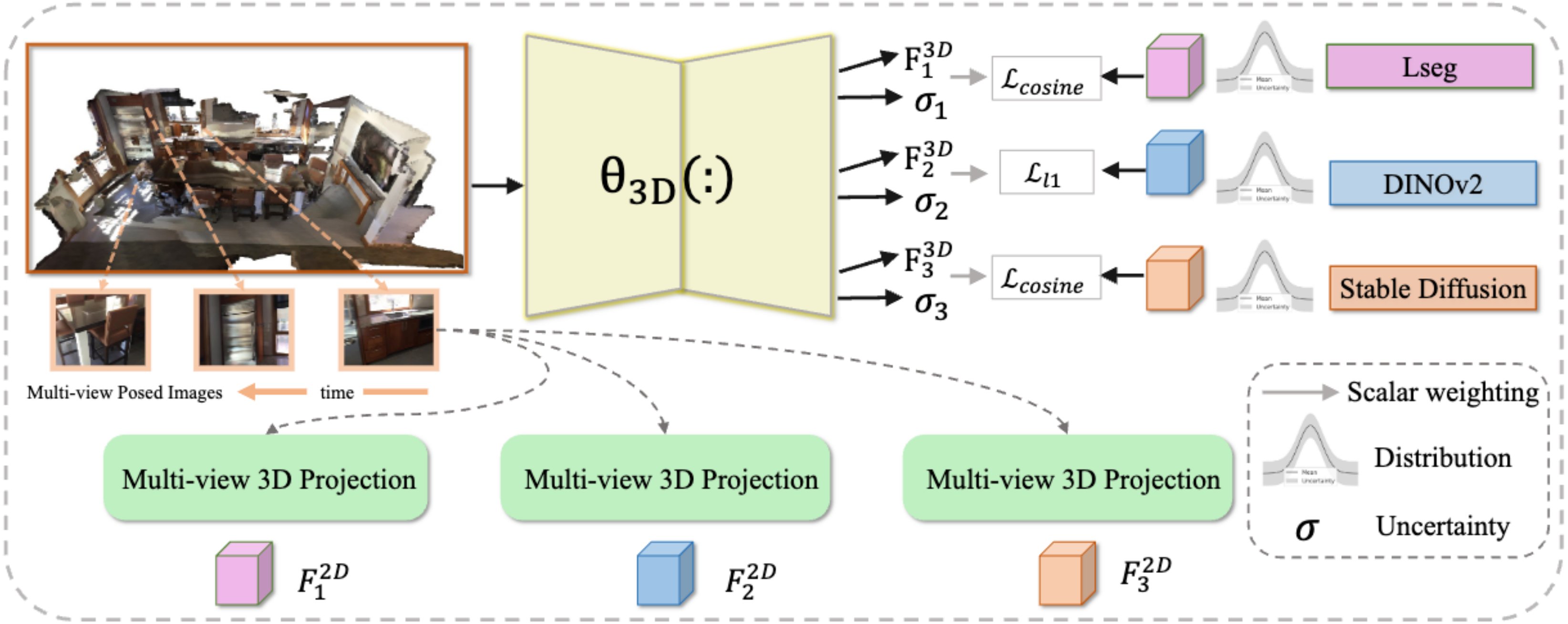
- Motivation
- Method
- Experiments

Method



- ➢ **Semantic priors & geometric knowledge** of spatially-aware 2D vision foundation models.
- ➢ **Deterministic uncertainty estimation** to capture ***uncertainties*** across diverse 2D embedding ambiguity.
- ➢ Helping with reconciling **heterogeneous representations** from 2D VLMs into one single 3D model.

Method



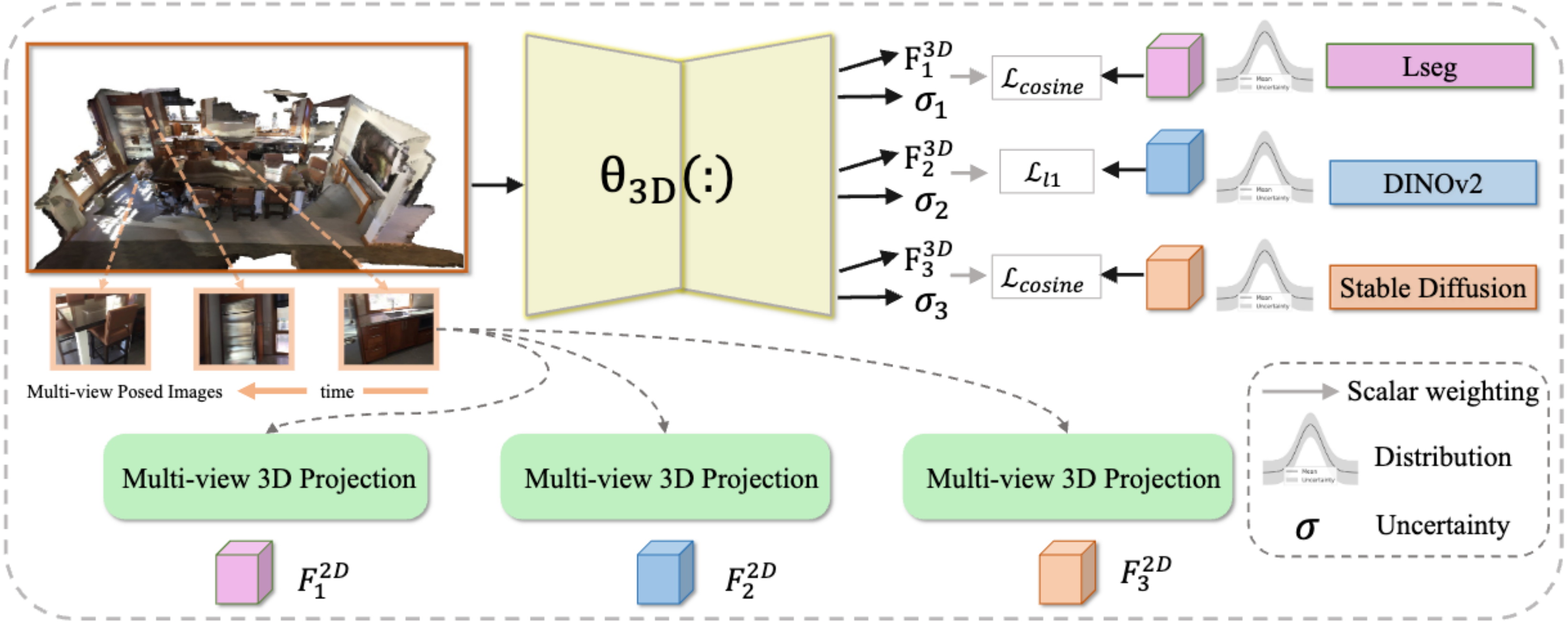
$$\mathcal{L}_{\cos\_lseg} = 1 - \frac{F_1^{3D} \cdot F_1^{2D}}{\left\|F_1^{3D}\right\|_2 \cdot \left\|F_1^{2D}\right\|_2}$$

$$\mathcal{L}_{l1\_DINOv2} = \frac{1}{n}\sum_{i=1}^{n}\left|F_2^{3D} - F_2^{2D}\right|$$

$$\mathcal{L}_{\cos\_sd} = 1 - \frac{F_3^{3D} \cdot F_3^{2D}}{\left\|F_3^{3D}\right\|_2 \cdot \left\|F_3^{2D}\right\|_2}, \qquad F_3^{2D} = F_3^{2D} - \mu_{F_3^{2D}}$$

$$\mathcal{L}_{distill} = \mathcal{L}_{\cos\_lseg} + \mathcal{L}_{l1\_DINOv2} + \mathcal{L}_{\cos\_sd}$$

Method



$$\mathcal{L}_{distill} = \mathcal{L}_{\cos\_lseg} + \mathcal{L}_{l1\_DINOv2} + \mathcal{L}_{\cos\_sd}$$

$$p(y_1, y_2, y_3 \mid f^W(x)) = \prod_{i=1}^{3} p(y_i \mid f^W(x) = \prod_{i=1}^{3} N(y_i;\ f^W(x), \sigma_i^2)$$

$$\mathcal{L}_{distill} = -\log p\big((y_1, y_2, y_3 \mid f^W(x)\big)$$

$$\propto \frac{1}{2\sigma_1^2}\mathcal{L}_{\cos\_lseg} + \frac{1}{2\sigma_2^2}\mathcal{L}_{l1\_DINOv2} + \frac{1}{2\sigma_3^2}\mathcal{L}_{\cos\_sd}$$

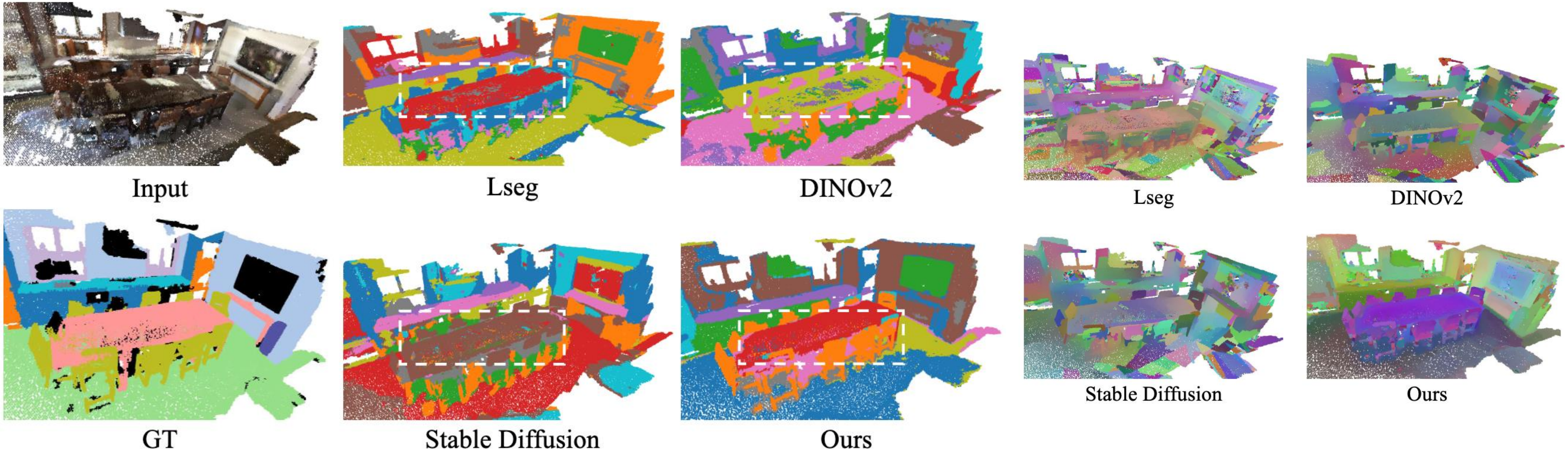
$$\log \sigma_i \ \rightarrow\ \log(1.0 +\ \sigma_i)$$

Outline

- Motivation
- Method
- Experiments

Experiments

2D feature structural visualizations from various VLMs --- K-Means UMAP

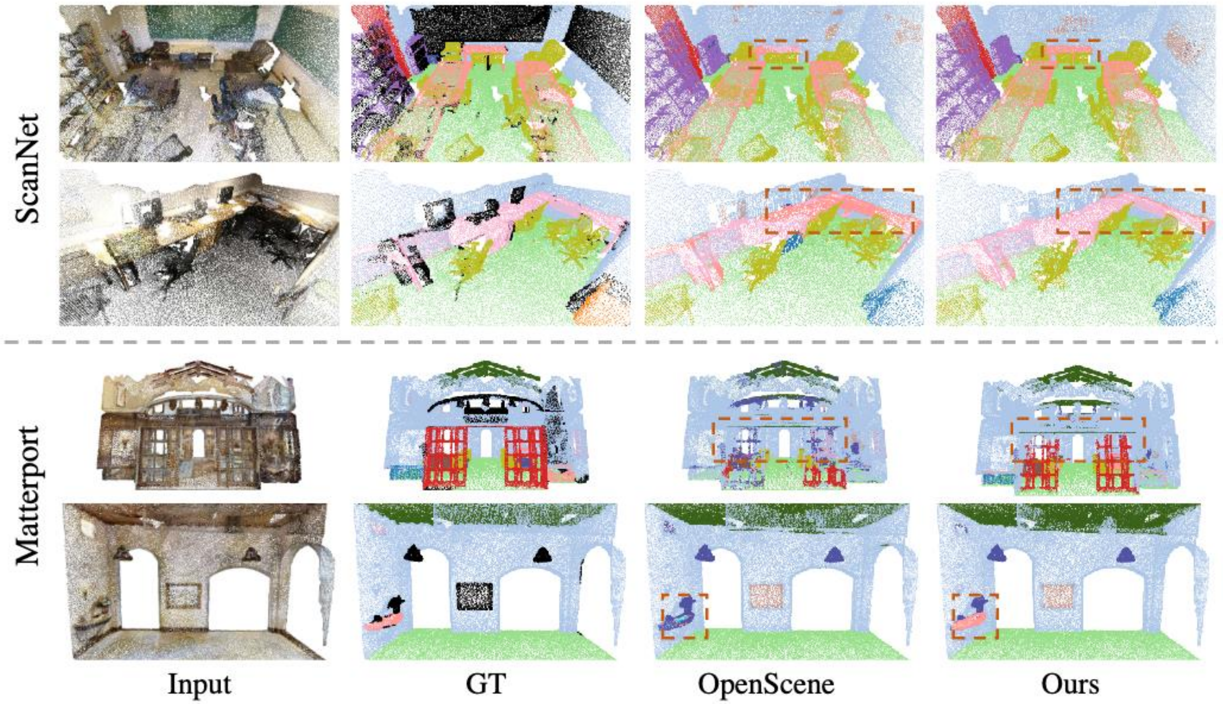


- **DINOv2** - smoother and more consistent results.
- **Diffusion Model** - intriguing geometric characteristics.
- **Lseg** – visual and text multi-modal alignment.

Experiments

Open-Vocabulary 3D Semantic Segmentation

Type	Method	ScanNetV2 mIoU	ScanNetV2 mAcc	Matterport3D mIoU	Matterport3D mAcc
Fully-sup.	TangentConv [80]	40.9	-	-	46.8
	TextureNet [31]	54.8	-	-	63.0
	ScanComplete [14]	56.6	-	-	44.9
	DCM-Net [77]	65.8	-	-	66.2
	Mix3D [58]	73.6	-	-	-
	SupCon [101]	69.2	77.7	53.1	63.4
	LGround [71]	73.2	-	-	67.2
	MinkowskiNet [11]	69.2	77.7	53.1	63.4
Upper-bound	MinkowskiNet^{reimple} [11]	68.96	77.41	54.12	65.57
Zero-shot	MSeg Voting [41]	45.6	54.4	33.4	-
	PLA [16]	17.7	33.5	-	-
	CLIP2Scene [9]	25.1	-	-	-
	CNS [10]	26.8	-	-	-
	CLIP-FO3D [94]	30.2	49.1	-	-
	RegionPLC [88]	43.8	65.6	-	-
	DMA-text only [46]	50.5	63.7	39.8	49.5
	OpenScene-3D† [64]	52.9	63.2	41.9	51.2
	OpenScene-2D3D† [64]	54.2	66.6	43.4	53.5
	OpenScene^{reimple}-3D [64]	51.6	63.1	40.5	48.8
	OpenScene^{reimple}-2D3D [64]	52.2	65.4	41.5	50.6
	(*Ours*) CUA-O3D (3D)	54.1	64.1	41.3	49.5
	(*Ours*) CUA-O3D (2D3D)	**55.3**	65.6	**42.2**	**50.9**



Open-vocabulary 3D semantic segmentation comparisons

Experiments

Cross-dataset evaluation

Method	mIoU	mAcc
ScanNetV2 (*train*) →Matterport3D (*eval*)		
OpenScene [64]	36.0	48.0
(*Ours*) CUA-O3D	**37.4** (+1.4)	**49.2**(+1.2)
Matterport3D (*train*)→ScanNetV2 (*eval*)		
OpenScene [64]	36.5	44.0
(*Ours*) CUA-O3D	**38.6** (+2.1)	**46.6**(+2.6)

Linear-probing segmentation

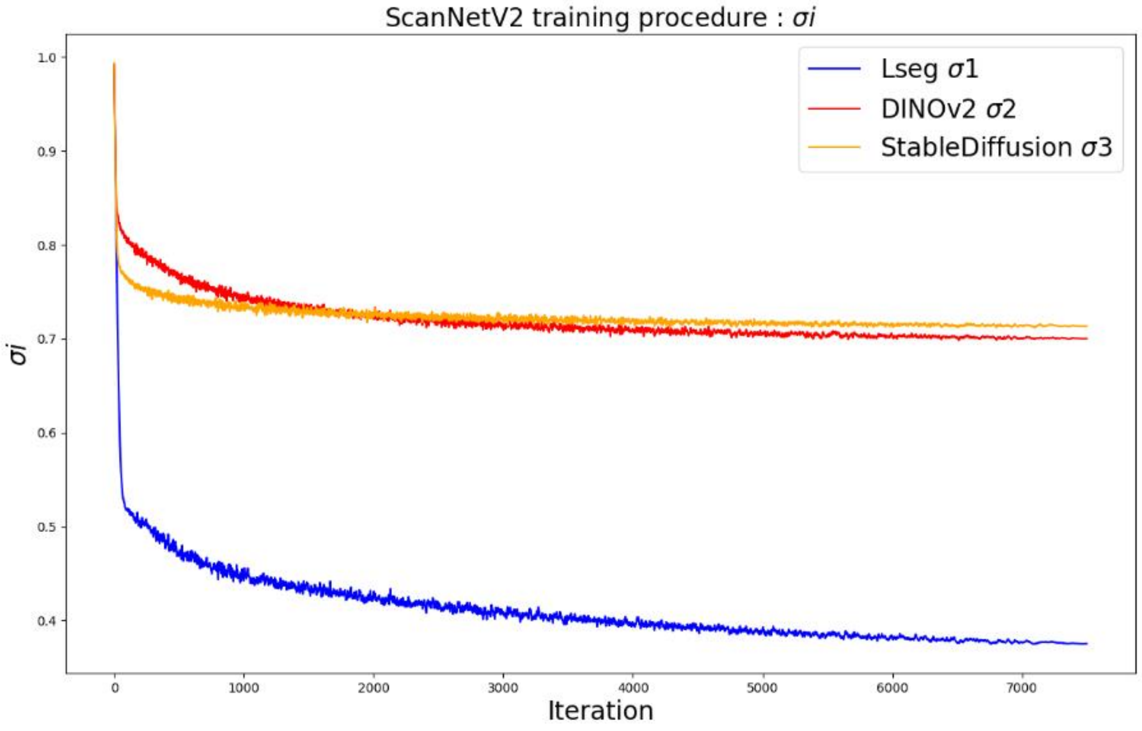
Type	Method	ScanNetV2		Matterport3D	
		mIoU	mAcc	mIoU	mAcc
Upperbound-fully sup.	MinkowskiNet [11]	68.9	77.4	54.1	65.5
Baseline init.	MinkowskiNet [11]	54.4	64.7	36.1	43.0
Concat	3-heads concat	62.1	72.7	45.8	55.3
Separate	3-heads average	61.7	72.0	45.4	55.0
Single-head	Lseg-head	59.9	71.5	-	-
	DINOv2-head	61.7	72.2	-	-
	StableDiffusion-head	61.4	72.1	-	-

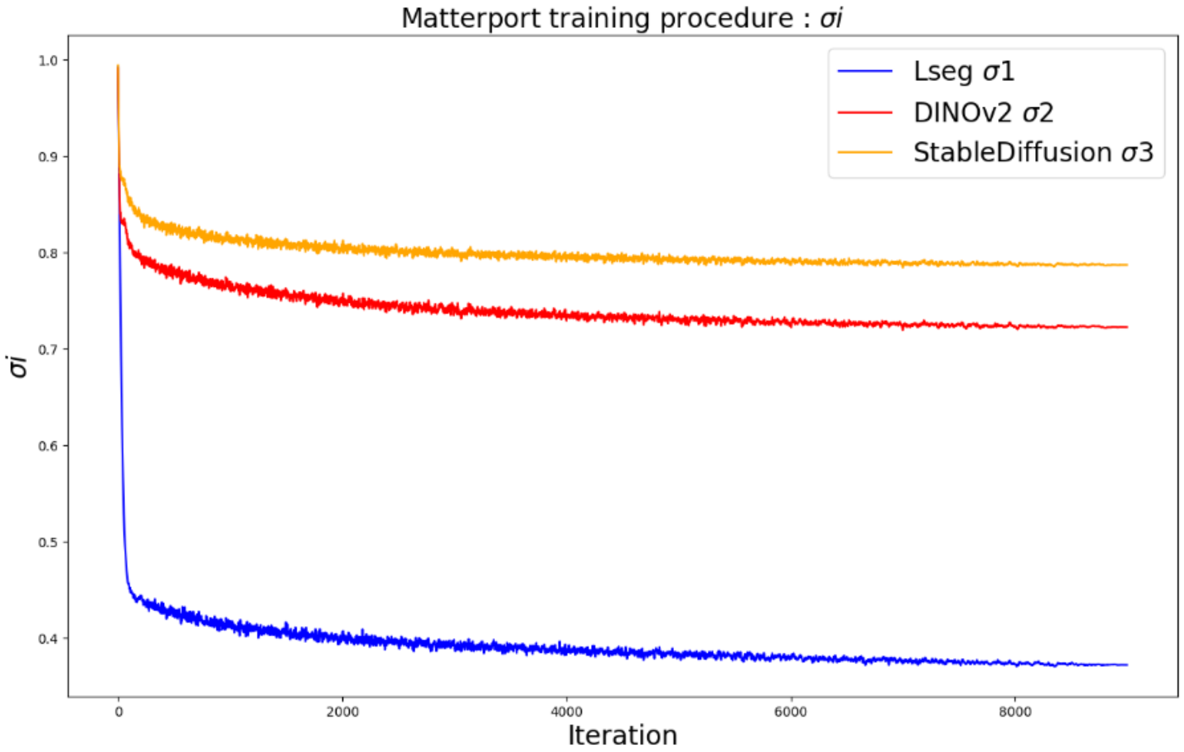
Cross-dataset generalization

(trained on ScanNetV2, and zero-shot tested on the Matterport3D)

Method	Matterport21		Matterport40		Matterport80		Matterport160	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
OpenScene‡ [64]	36.0	48.0	21.1	27.5	10.8	13.9	6.0	8.1
(*Ours*) CUA-O3D (2D3D)	**37.4**	**49.2**	**23.3**	**30.2**	**12.2**	**16.3**	**6.1**	**8.4**

Experiments





Evolutions of parameters in terms of deterministic uncertainty estimation σ_i.

Project Pages

Codes