NEURAL INFORMATION PROCESSING SYSTEMS

Diffusion-SS3D: Diffusion Model for Semi-supervised 3D Object Detection

Cheng-Ju Ho[1] **Chen-Hsuan Tai**[1] **Yen-Yu Lin**[1] **Ming-Hsuan Yang**[2,3] **Yi-Hsuan Tsai**[3]

[1]National Yang Ming Chiao Tung University

[2]University of California at Merced

[3]Google

Semi-supervised Learning (SSL)

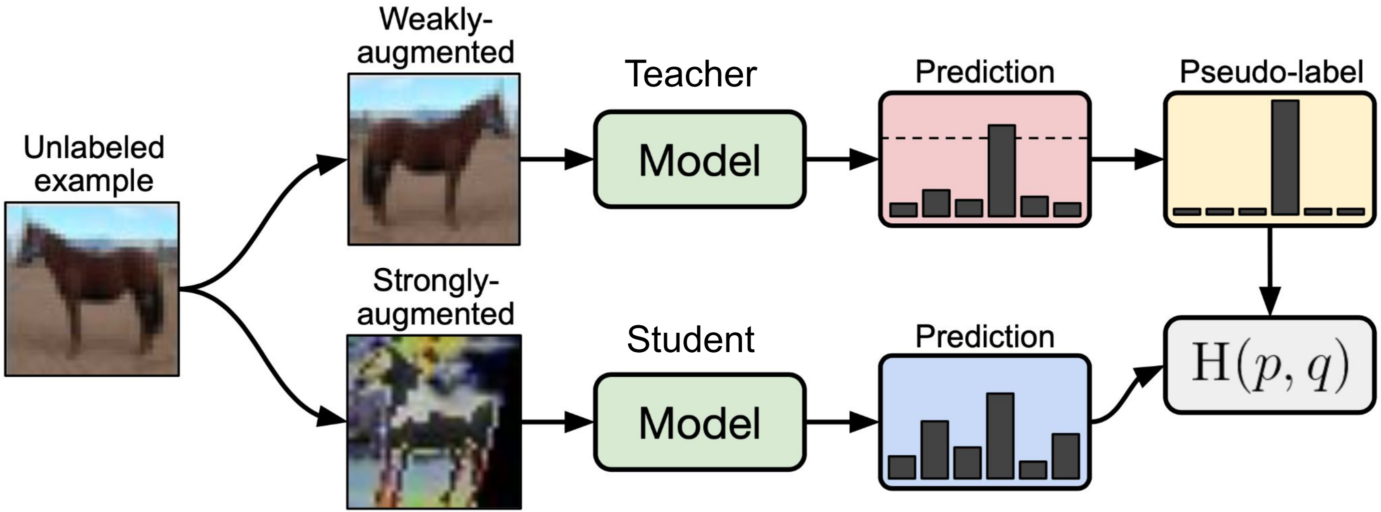
- The teacher-student framework utilizes pseudo-labels as supervisory signals for unlabeled data.

Unlabeled example

Semi-supervised learning methods incorporate unlabeled data in the model training process.

Semi-supervised Learning (SSL)

- The teacher-student framework utilizes pseudo-labels as supervisory signals for unlabeled data.

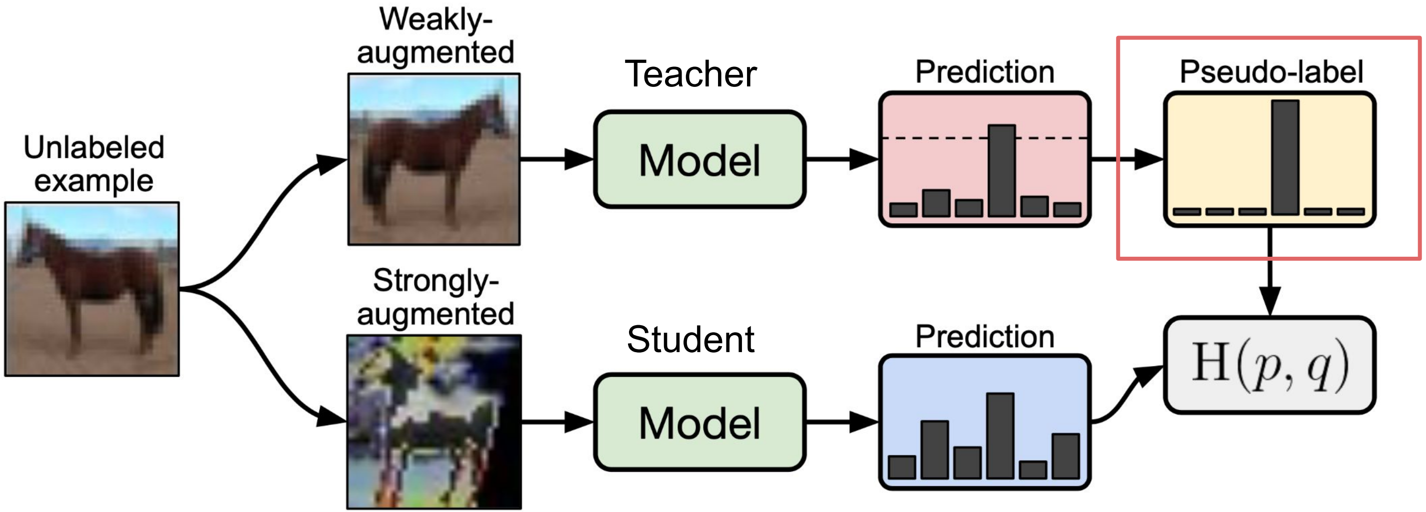


[FixMatch, Sohn et al. NIPS'20]

Among those methods, the teacher-student framework is a widely used approach, which utilizes pseudo-labels as supervisory signals for unlabeled data.

Semi-supervised Learning (SSL)

- The teacher-student framework utilizes pseudo-labels as supervisory signals for unlabeled data.



FixMatch, Sohn et al. NIPS'20

Among those methods, the teacher-student framework is a widely used approach, which utilizes pseudo-labels as supervisory signals for unlabeled data.

Motivation

- Challenges in SSL 3D object detection

However, current SSL 3D object detection methods encounter some challenges.

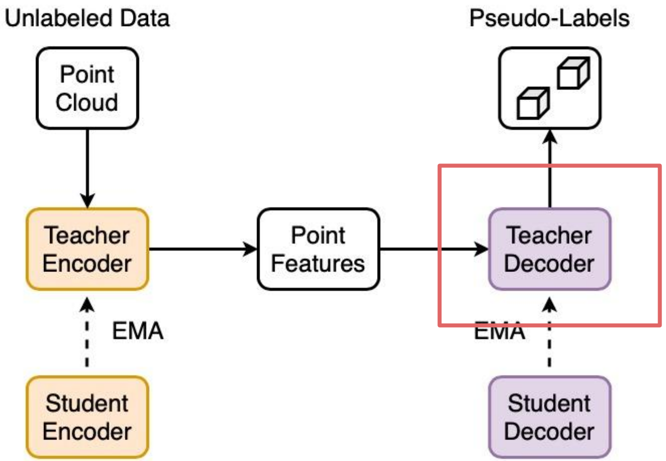
Motivation

- Challenges in SSL 3D object detection
  - The complexity of generating high-quality pseudo-labels.

For example, generating high-quality pseudo-labels is difficult due to diverse object locations in 3D space.

Motivation

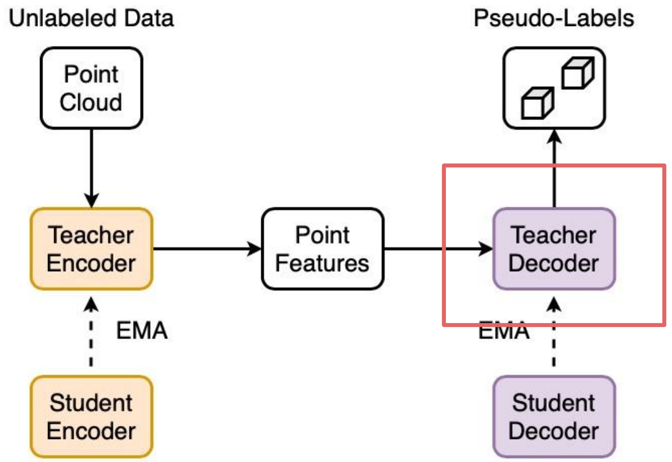
- Challenges in SSL 3D object detection
  - The complexity of generating high-quality pseudo-labels.
  - The model's outputs are the only sources to generate candidates for pseudo-labels.



Also, the model's outputs are the only source to generate pseudo-label candidates.

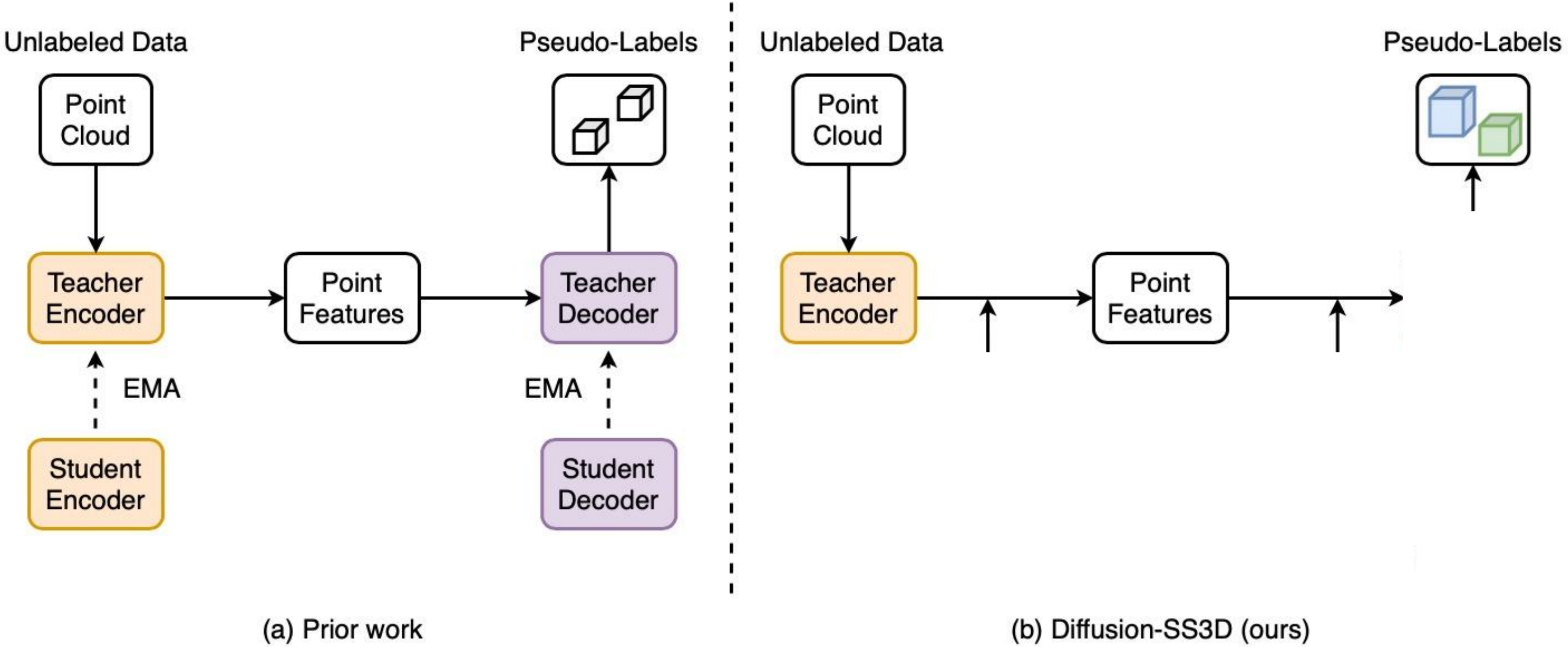
Motivation

- Challenges in SSL 3D object detection
  - The complexity of generating high-quality pseudo-labels.
  - The model's outputs are the only sources to generate candidates for pseudo-labels.



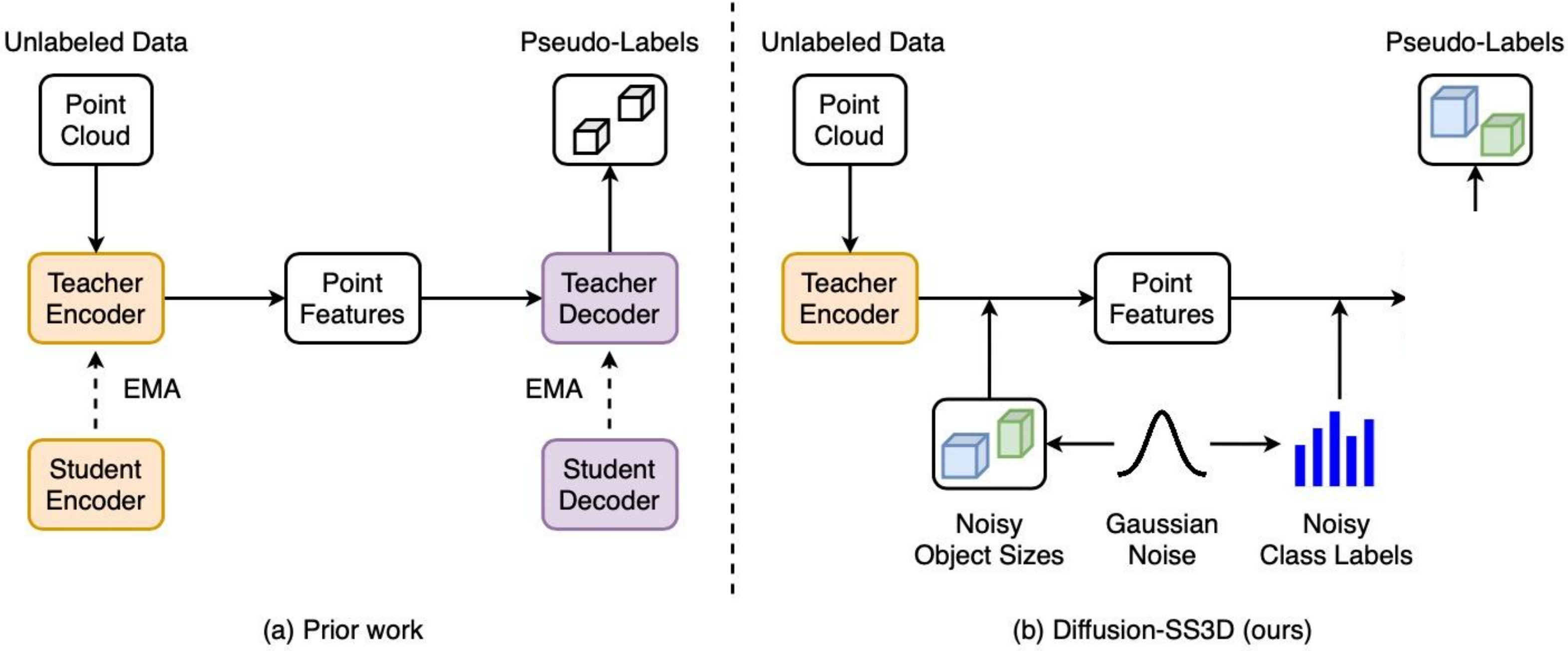
That poses issues when objects remain undetected due to insufficient predictions, resulting in a lower recall rate.

Overview of Diffusion-SS3D



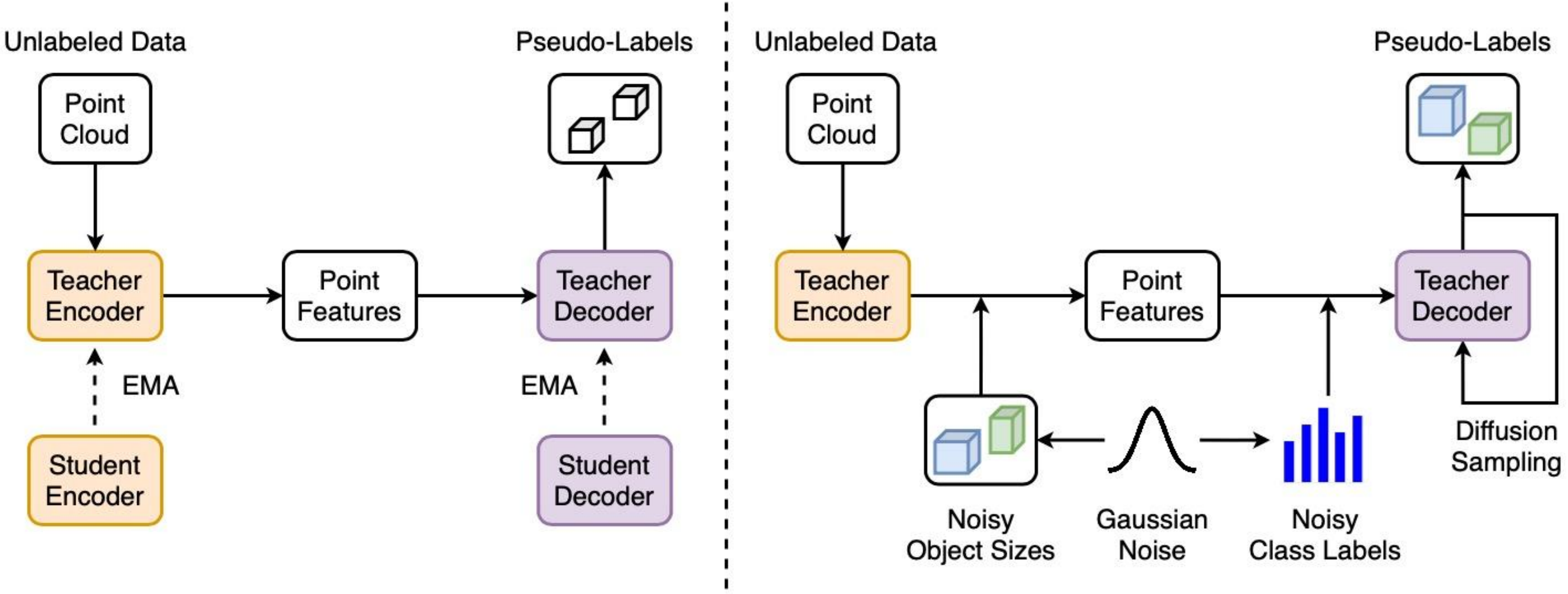
Instead, Diffusion-SS3D improves the quality of pseudo-labels through a diffusion model.

Overview of Diffusion-SS3D



We include random noises to produce corrupted 3D object size and class label distributions

Overview of Diffusion-SS3D



(a) Prior work

(b) Diffusion-SS3D (ours)

and utilize an iterative denoising process to generate reliable pseudo-labels.

Extract RoI-features with Noisy Boxes

Here, we focus on extracting RoI-features from noisy bounding boxes.

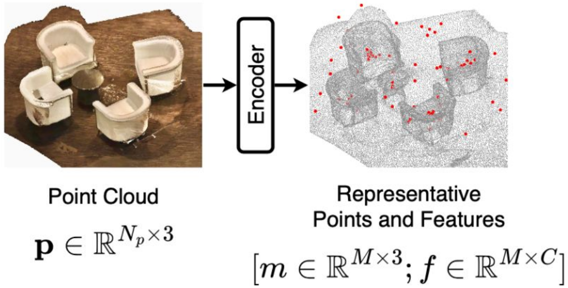
Extract RoI-features with Noisy Boxes

Point Cloud

$\mathbf{p} \in \mathbb{R}^{N_p \times 3}$

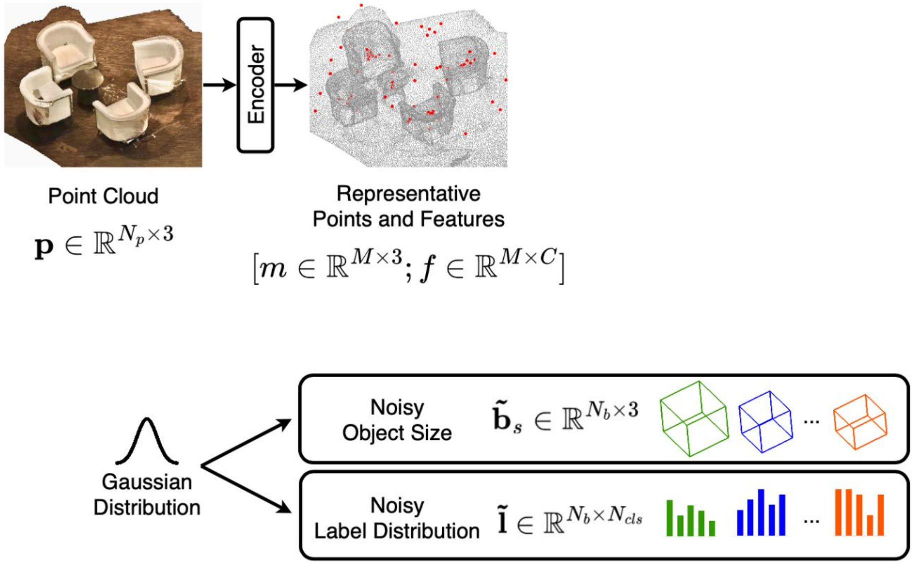
First, we start with an input point cloud scene containing N_P points.

Extract RoI-features with Noisy Boxes



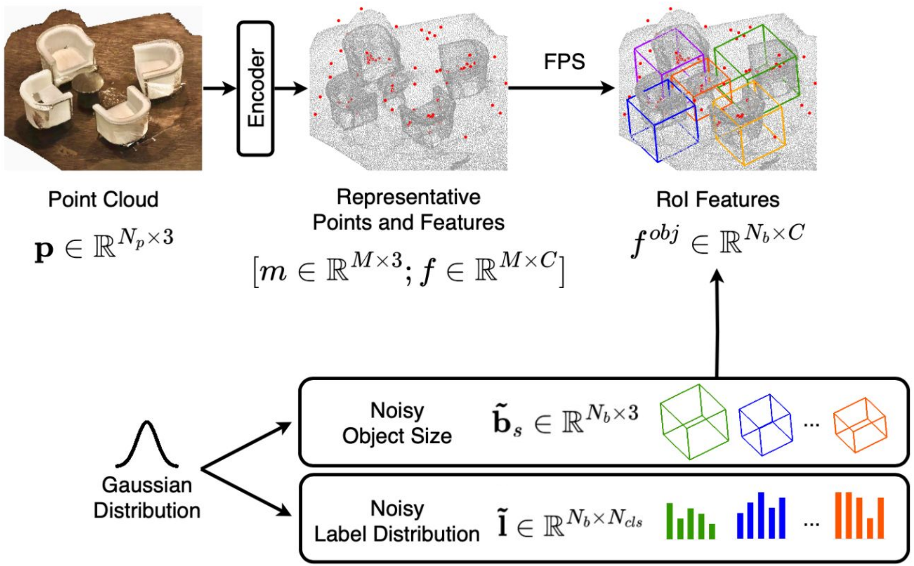
The encoder then extracts representative features, where each feature consists of a potential object center *m* and its corresponding high-level feature *f*.

Extract RoI-features with Noisy Boxes



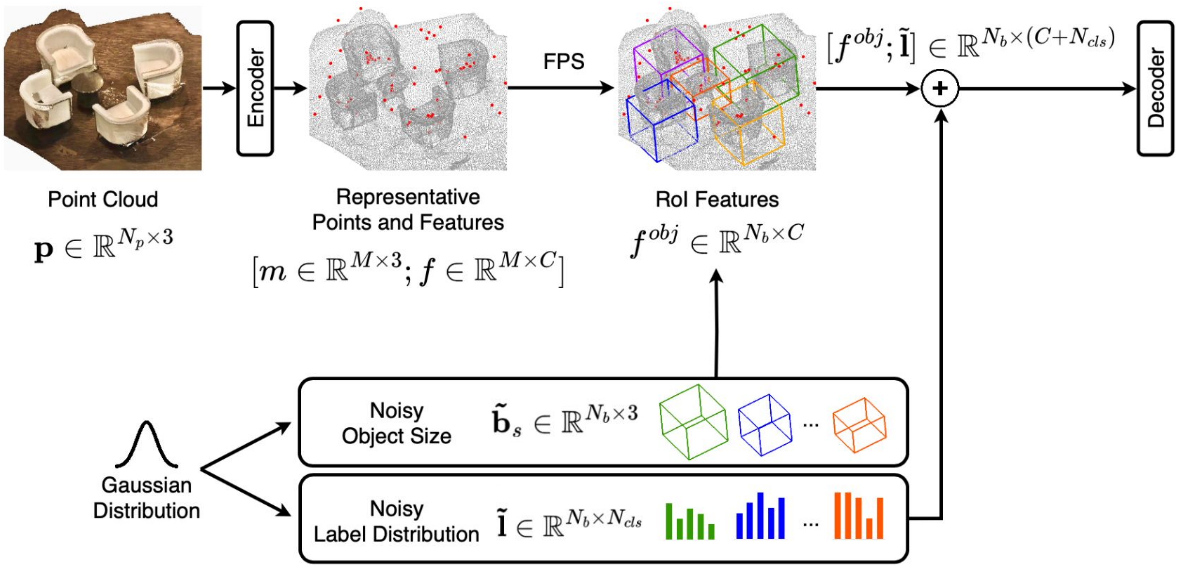
We sample N_b noisy object sizes and noisy label distributions from Gaussian noise.

Extract RoI-features with Noisy Boxes



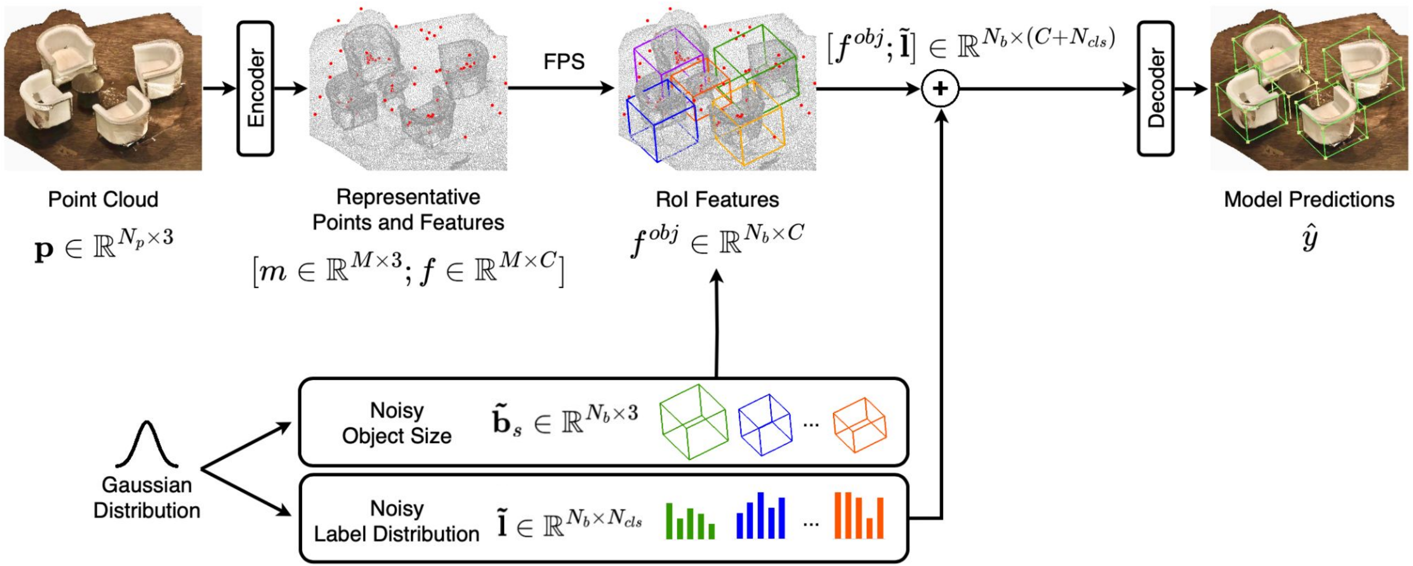
Following farthest point sampling, the N_b representative points serve as box centers, combining with noisy sizes to form noisy boxes for gathering RoI-features.

Extract RoI-features with Noisy Boxes



Subsequently, we merge the noisy label distributions with RoI-features and input them into the decoder.

Extract RoI-features with Noisy Boxes

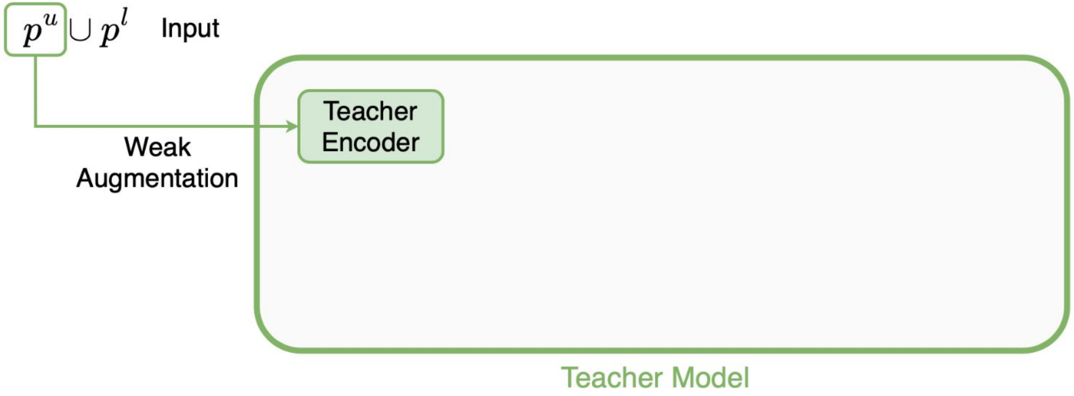


Our decoder is trained to make accurate predictions, even when dealing with RoI-features extracted from noisy boxes and their associated noisy label conditions.

Diffusion-SS3D Framework for SSL

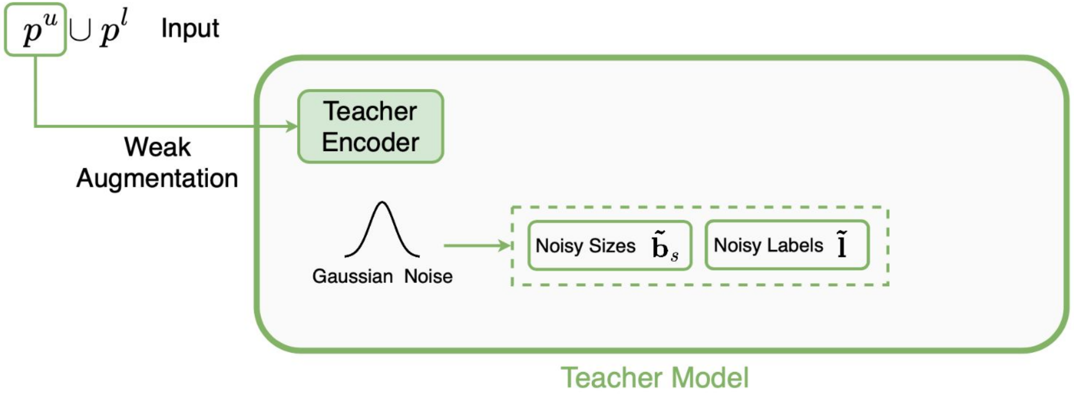
Let's now demonstrate how our method integrates into the teacher-student framework.

Diffusion-SS3D Framework for SSL



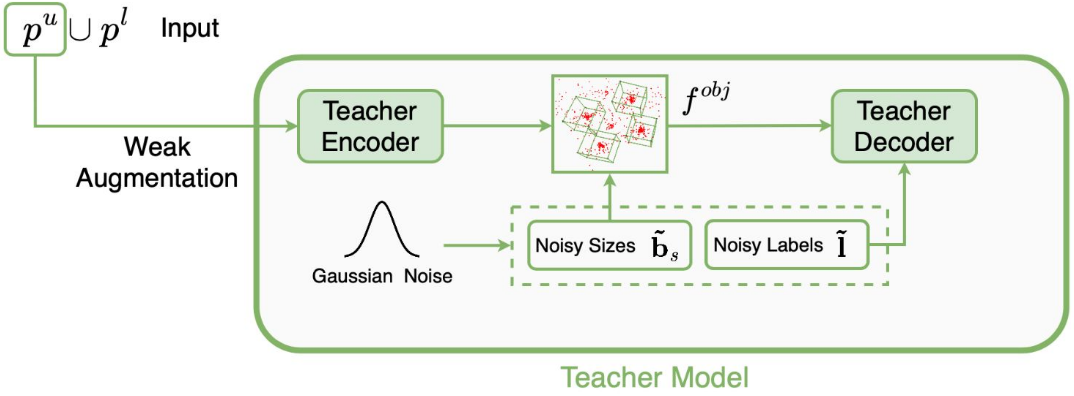
First, unlabeled data undergo weak augmentation and are fed into the teacher model, producing pseudo-labels.

Diffusion-SS3D Framework for SSL



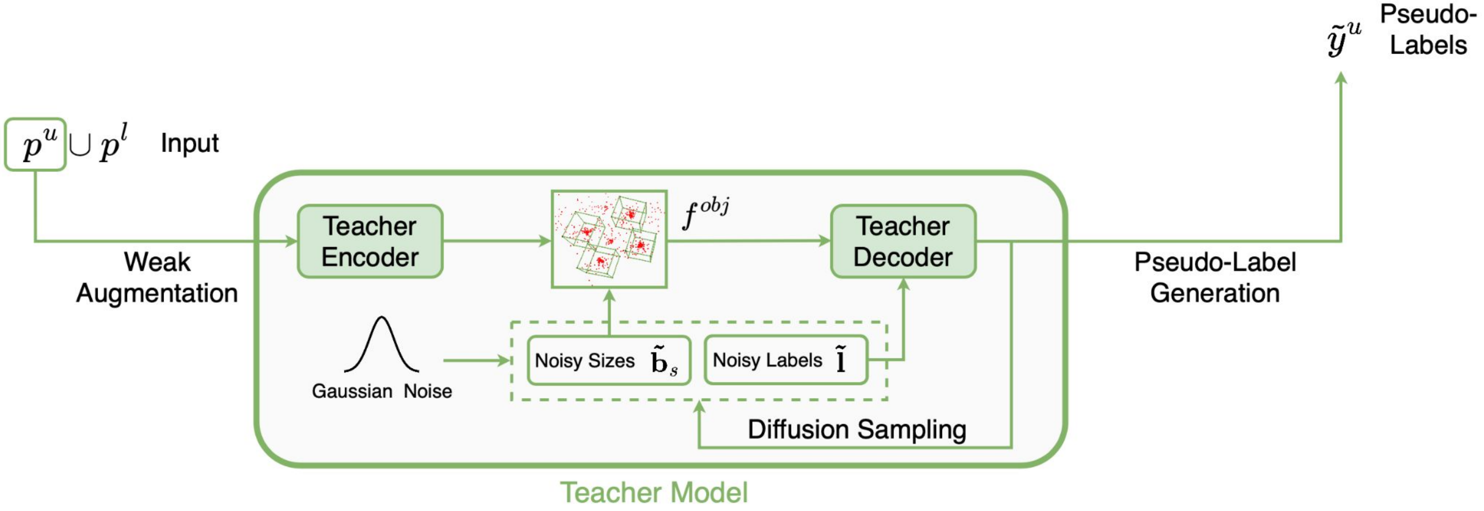
Concurrently, we generate noisy sizes and label distributions from Gaussian noise.

Diffusion-SS3D Framework for SSL



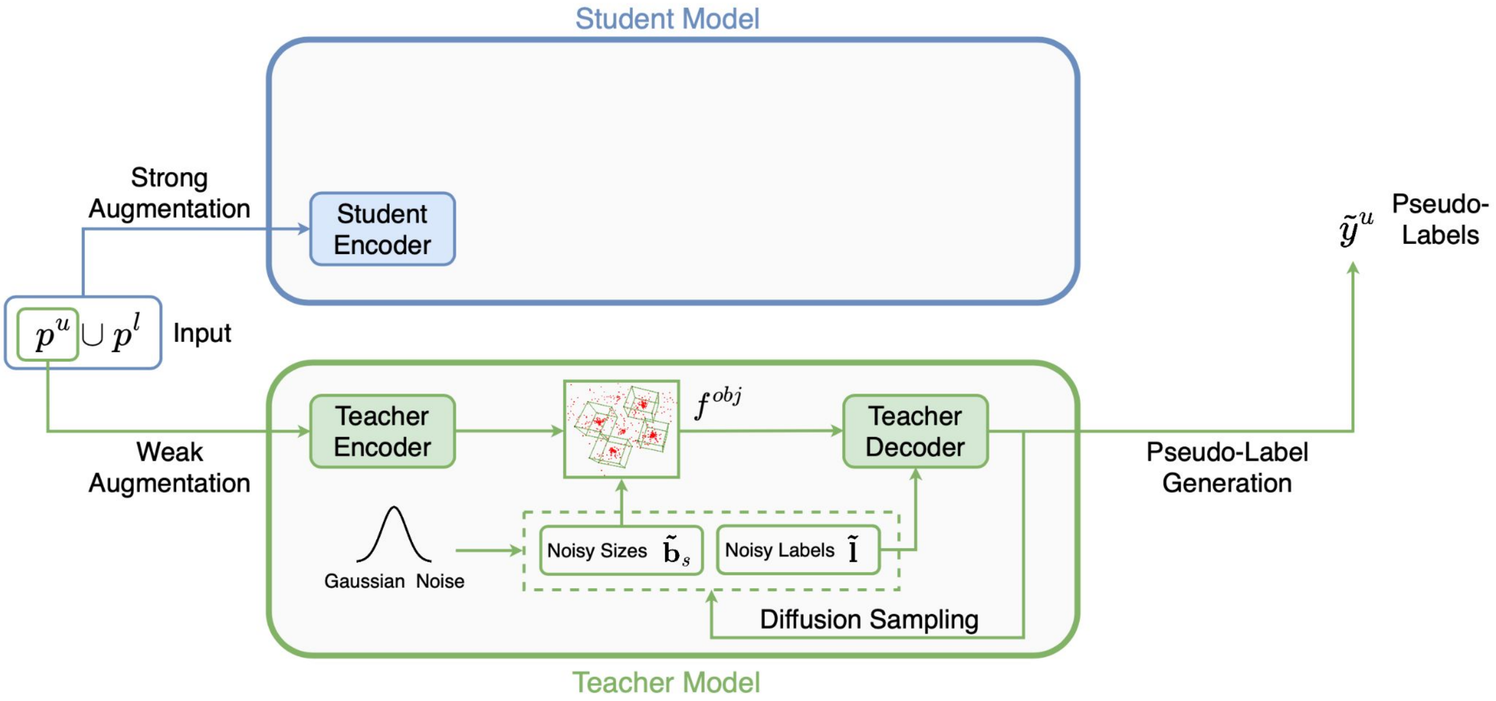
As explained earlier, these noisy features and label information are integrated into the teacher decoder, enabling it to make initial predictions.

Diffusion-SS3D Framework for SSL



We then refine these predictions through iterative diffusion sampling, yielding high-quality pseudo-labels.

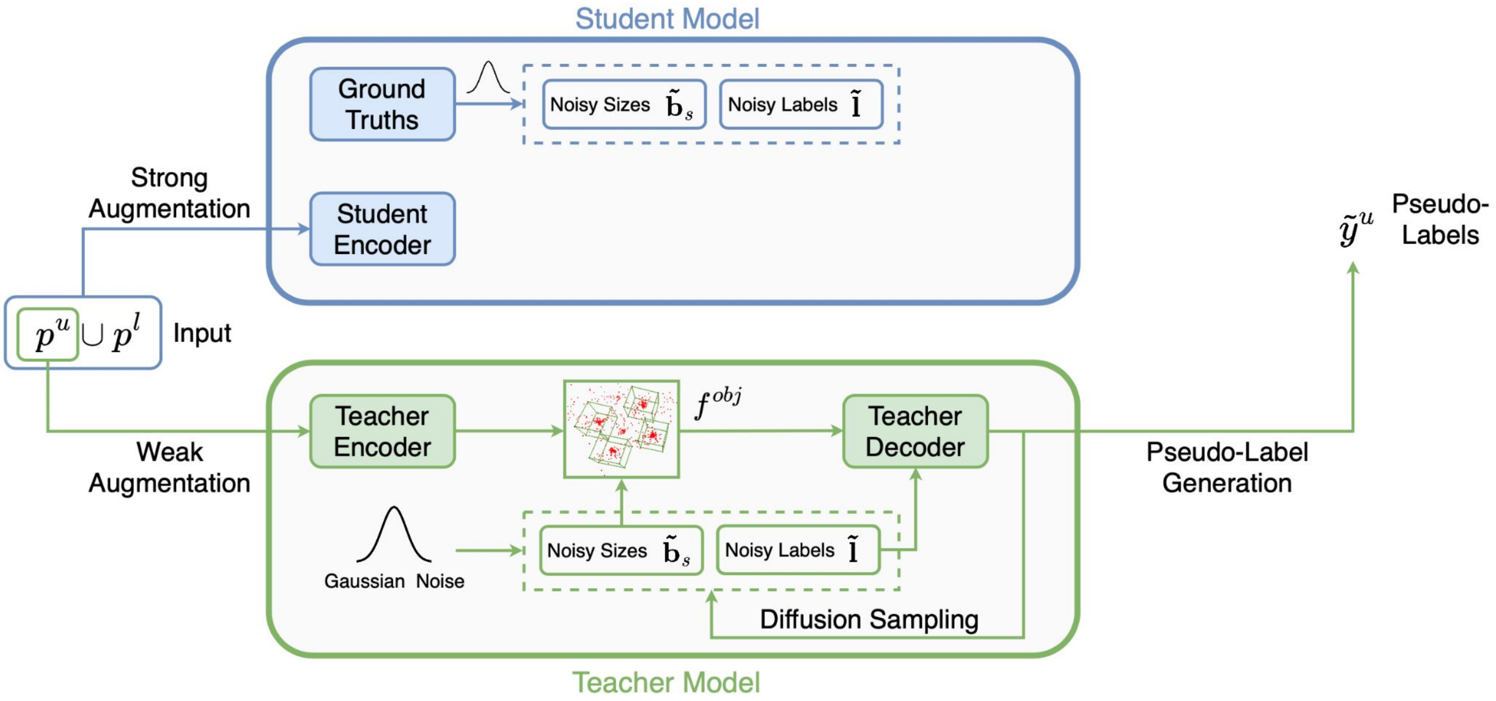
Diffusion-SS3D Framework for SSL



In contrast, both labeled and unlabeled data undergo strong augmentation and are fed into the student model for training.

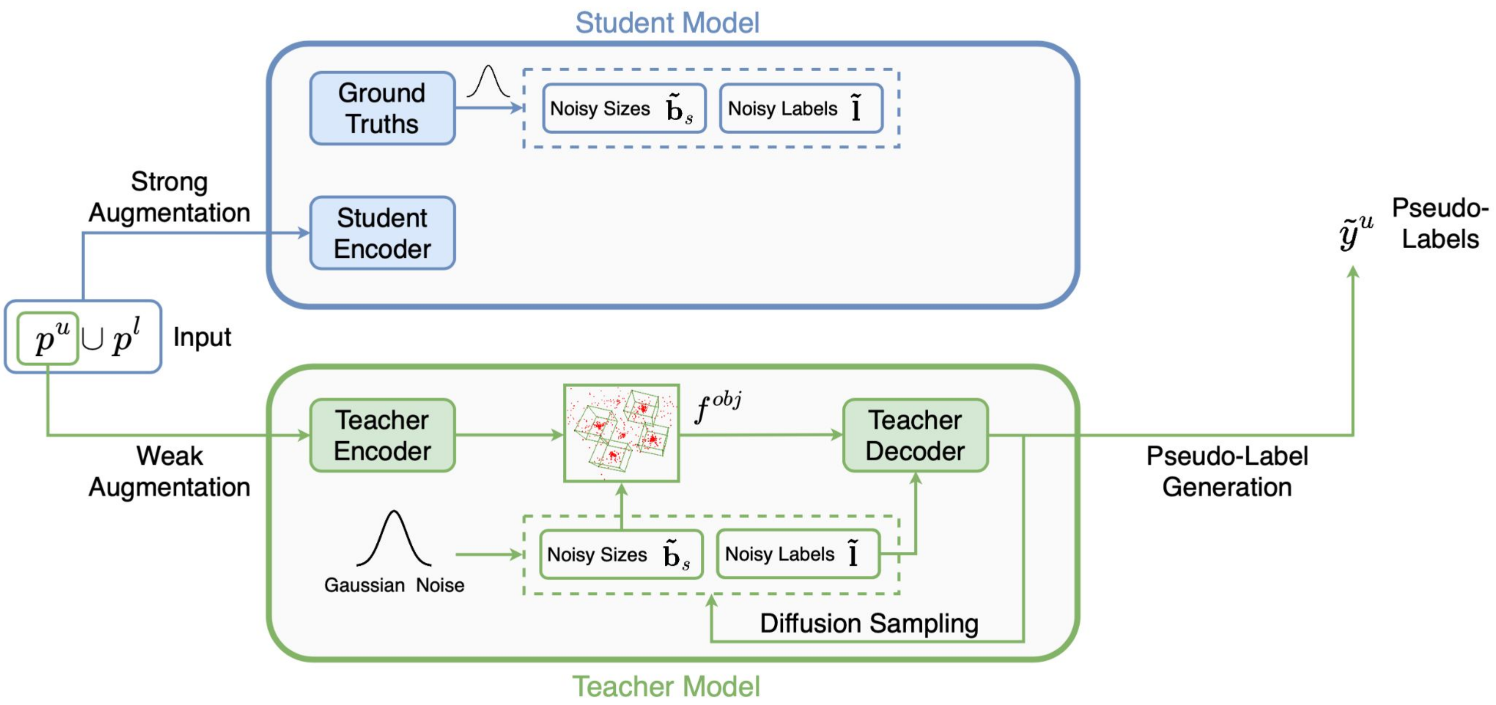
Diffusion-SS3D Framework for SSL

NEURAL INFORMATION PROCESSING SYSTEMS



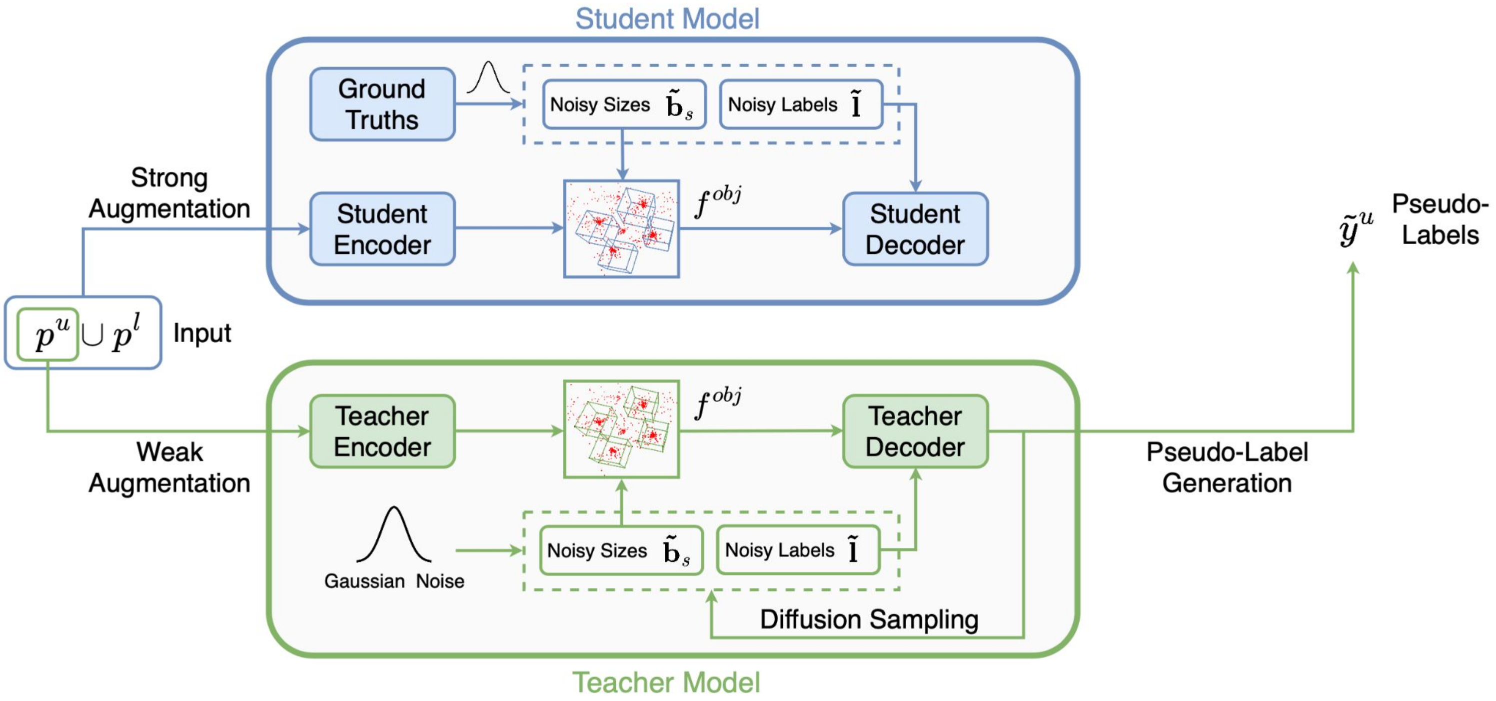
Gaussian noise is introduced to ground-truth sizes and class labels to create noisy samples.

Diffusion-SS3D Framework for SSL



Instead, pseudo-labels are utilized for unlabeled data.

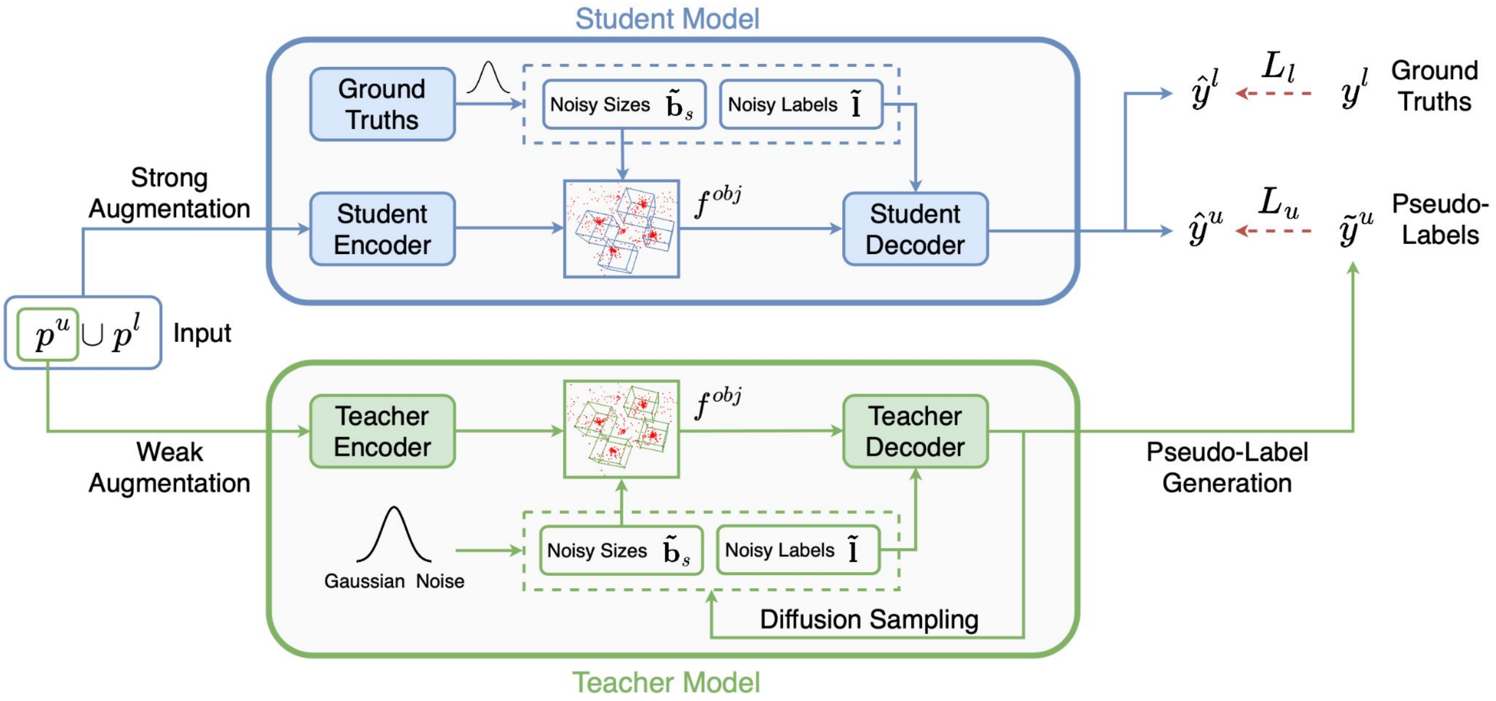
Diffusion-SS3D Framework for SSL



The student decoder takes these noisy features as input, refining them in a single step to make accurate predictions.

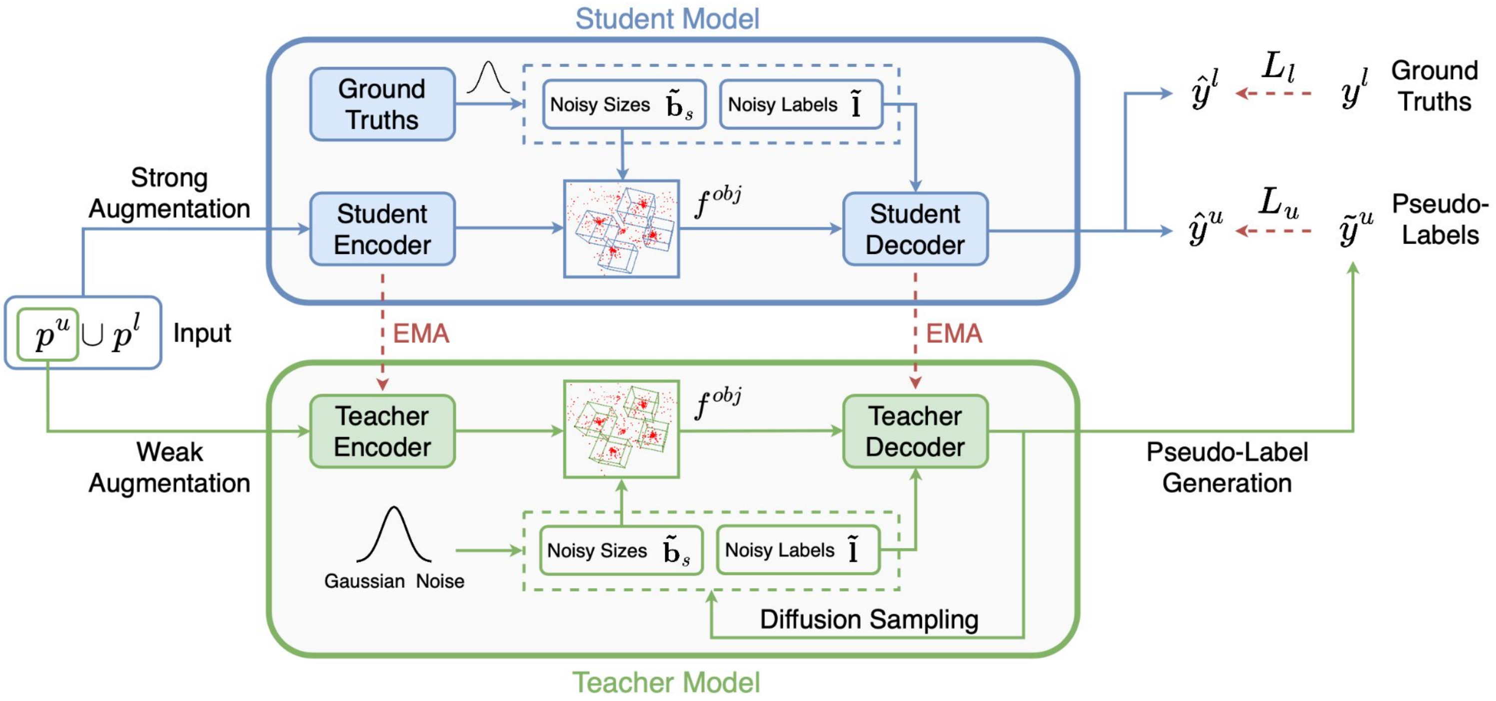
Diffusion-SS3D Framework for SSL

NEURAL INFORMATION PROCESSING SYSTEMS



Subsequently, we calculate the detection loss to update the student model.

Diffusion-SS3D Framework for SSL



Finally, the teacher model is updated from the student model through the EMA mechanism.

Experimental Results

Table 1: Results on the ScanNet val set with 5%, 10%, 20%, and 100% labeled data.

Model	5%		10%		20%		100%	
	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5
VoteNet	27.9 ± 0.5	10.8 ± 0.6	36.9 ± 1.6	18.2 ± 1.0	46.9 ± 1.9	27.5 ± 1.2	57.8	36.0
SESS	32.0 ± 0.7	14.4 ± 0.7	39.5 ± 1.8	19.8 ± 1.3	49.6 ± 1.1	29.0 ± 1.0	61.3	38.8
3DIoUMatch	40.0 ± 0.9	22.5 ± 0.5	47.2 ± 0.4	28.3 ± 1.5	52.8 ± 1.2	35.2 ± 1.1	62.9	42.1
Diffusion-SS3D	**43.5** ± 0.2	**27.9** ± 0.3	**50.3** ± 1.4	**33.1** ± 1.5	**55.6** ± 1.7	**36.9** ± 1.4	**64.1**	**43.2**
Gain (mAP)	3.5↑	5.4↑	3.1↑	4.8↑	2.8↑	1.7↑	1.2↑	1.1↑

Table 2: Results on the SUN RGB-D val set with 1%, 5% 10%, and 20% labeled data.

Model	1%		5%		10%		20%	
	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5
VoteNet	18.3 ± 1.2	4.4 ± 0.4	29.9 ± 1.5	10.5 ± 0.5	38.9 ± 0.8	17.2 ± 1.3	45.7 ± 0.6	22.5 ± 0.8
SESS	20.1 ± 0.2	5.8 ± 0.3	34.2 ± 2.0	13.1 ± 1.0	42.1 ± 1.1	20.9 ± 0.3	47.1 ± 0.7	24.5 ± 1.2
3DIoUMatch	21.9 ± 1.4	8.0 ± 1.5	39.0 ± 1.9	21.1 ± 1.7	45.5 ± 1.5	28.8 ± 0.7	49.7 ± 0.4	30.9 ± 0.2
Diffusion-SS3D	**30.9** ± 1.0	**14.7** ± 1.2	**43.9** ± 0.6	**24.9** ± 0.3	**49.1** ± 0.5	**30.4** ± 0.7	**51.4** ± 0.8	**32.4** ± 0.6
Gain (mAP)	9.0↑	6.7↑	4.9↑	3.8↑	3.6↑	1.6↑	1.7↑	1.5↑

We report the results of Diffusion-SS3D on the ScanNet and SUN RGB-D datasets with different amounts of labeled data.

Experimental Results

Table 1: Results on the ScanNet val set with 5%, 10%, 20%, and 100% labeled data.

Model	5%		10%		20%		100%	
	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5
VoteNet	27.9 ± 0.5	10.8 ± 0.6	36.9 ± 1.6	18.2 ± 1.0	46.9 ± 1.9	27.5 ± 1.2	57.8	36.0
SESS	32.0 ± 0.7	14.4 ± 0.7	39.5 ± 1.8	19.8 ± 1.3	49.6 ± 1.1	29.0 ± 1.0	61.3	38.8
3DIoUMatch	40.0 ± 0.9	22.5 ± 0.5	47.2 ± 0.4	28.3 ± 1.5	52.8 ± 1.2	35.2 ± 1.1	62.9	42.1
Diffusion-SS3D	**43.5** ± 0.2	**27.9** ± 0.3	**50.3** ± 1.4	**33.1** ± 1.5	**55.6** ± 1.7	**36.9** ± 1.4	**64.1**	**43.2**
Gain (mAP)	3.5↑	5.4↑	3.1↑	4.8↑	2.8↑	1.7↑	1.2↑	1.1↑

Table 2: Results on the SUN RGB-D val set with 1%, 5% 10%, and 20% labeled data.

Model	1%		5%		10%		20%	
	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5	mAP @ 0.25	mAP @ 0.5
VoteNet	18.3 ± 1.2	4.4 ± 0.4	29.9 ± 1.5	10.5 ± 0.5	38.9 ± 0.8	17.2 ± 1.3	45.7 ± 0.6	22.5 ± 0.8
SESS	20.1 ± 0.2	5.8 ± 0.3	34.2 ± 2.0	13.1 ± 1.0	42.1 ± 1.1	20.9 ± 0.3	47.1 ± 0.7	24.5 ± 1.2
3DIoUMatch	21.9 ± 1.4	8.0 ± 1.5	39.0 ± 1.9	21.1 ± 1.7	45.5 ± 1.5	28.8 ± 0.7	49.7 ± 0.4	30.9 ± 0.2
Diffusion-SS3D	**30.9** ± 1.0	**14.7** ± 1.2	**43.9** ± 0.6	**24.9** ± 0.3	**49.1** ± 0.5	**30.4** ± 0.7	**51.4** ± 0.8	**32.4** ± 0.6
Gain (mAP)	9.0↑	6.7↑	4.9↑	3.8↑	3.6↑	1.6↑	1.7↑	1.5↑

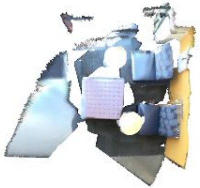
Overall, our method performs favorably against state-of-the-art approaches.

Visualization

Here, we visualize how diffusion-SS3D denoise via DDIM sampling step during inference.

Visualization

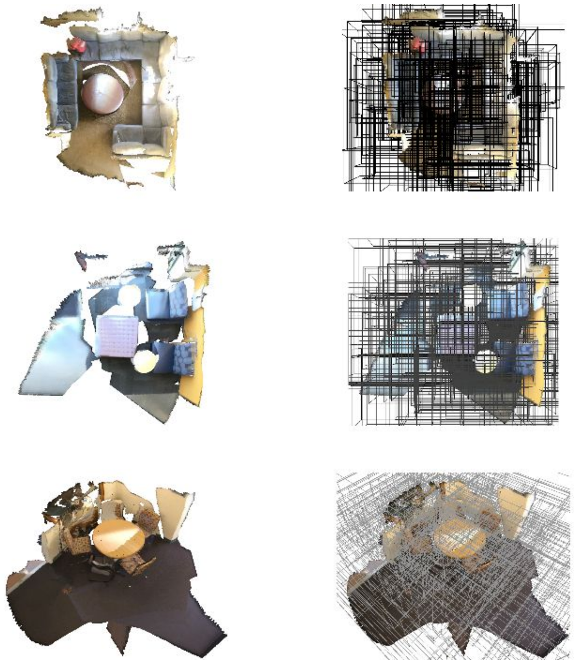
(a) Input point cloud



In each example, we show (a) the input point cloud.

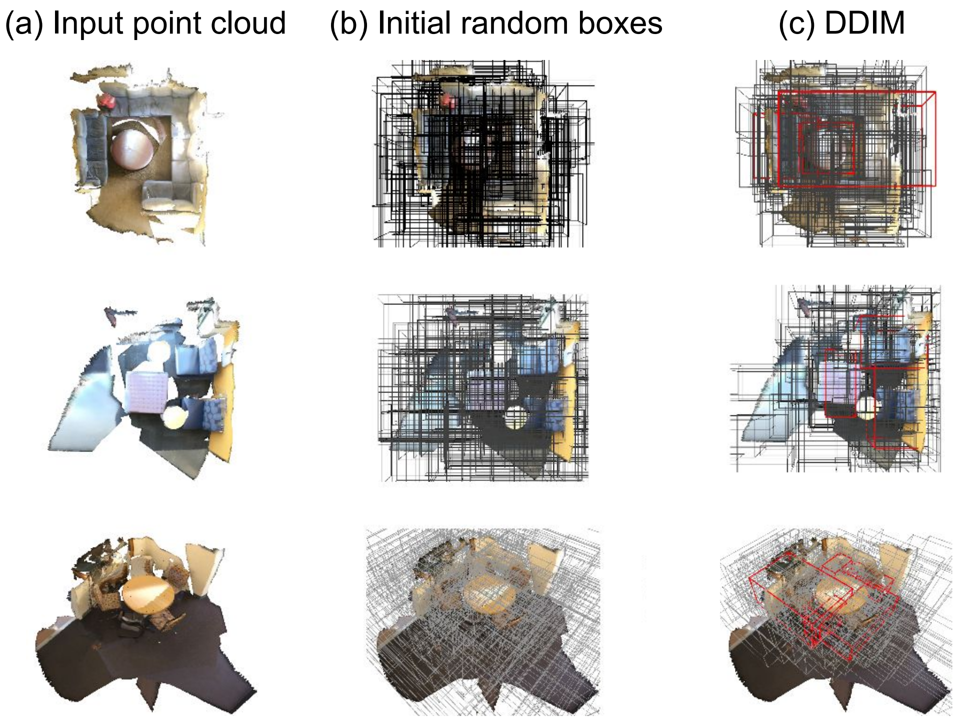
Visualization

(a) Input point cloud (b) Initial random boxes



(b) The initial bounding boxes obtained by random sampling.

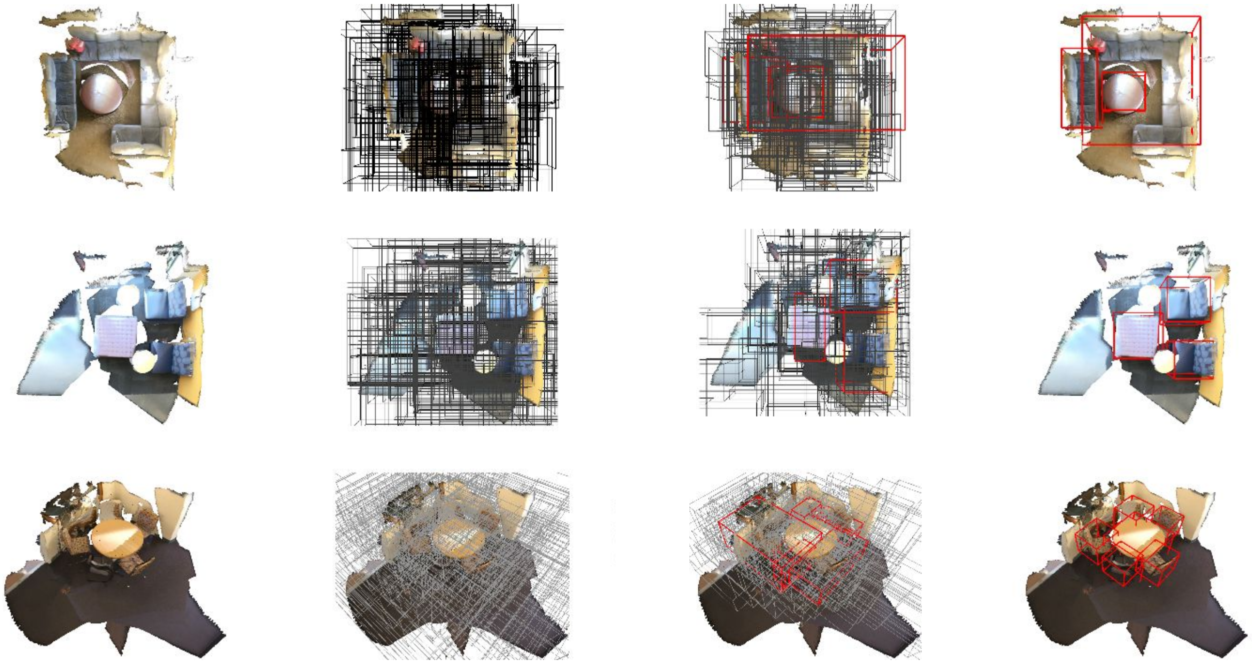
Visualization



(c) The denoised bounding boxes yielded by DDIM, where those closest to the ground truth are highlighted in red.

Visualization

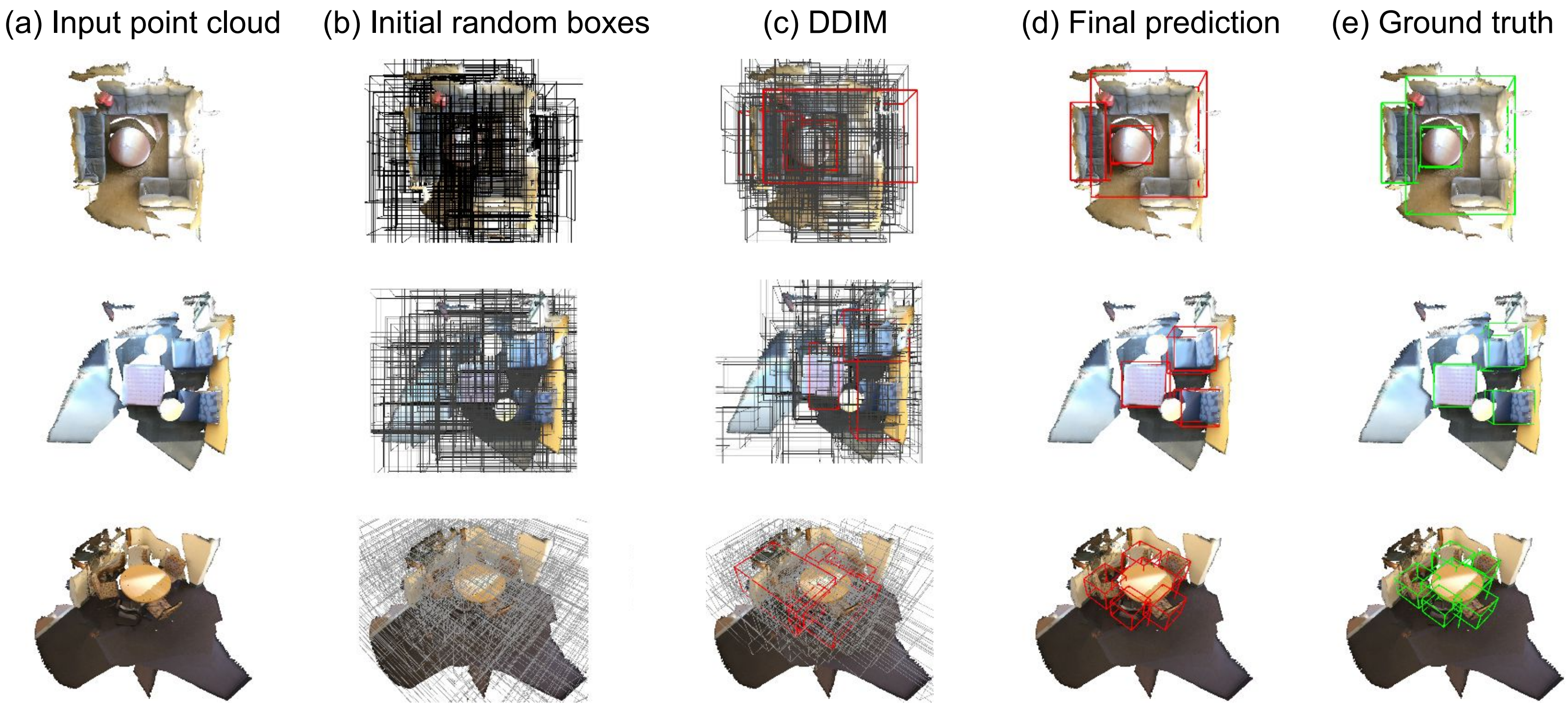
(a) Input point cloud (b) Initial random boxes (c) DDIM (d) Final prediction



(d) The detection results given by our diffusion decoder.

NEURAL INFORMATION PROCESSING SYSTEMS

Visualization



and (e) the ground-truth bounding boxes.

Conclusion

- We are the first method to utilize the diffusion model for SSL 3D object detection, treating the task as a denoising process for improving the quality of pseudo-labels.

Conclusion

- We are the first method to utilize the diffusion model for SSL 3D object detection, treating the task as a denoising process for improving the quality of pseudo-labels.

- We introduce the random noise to 3D object size and class label distributions for producing more plausible pseudo bounding boxes, by a means to integrate the diffusion model into the teacher-student framework for SSL.

Conclusion

- We are the first method to utilize the diffusion model for SSL 3D object detection, treating the task as a denoising process for improving the quality of pseudo-labels.

- We introduce the random noise to 3D object size and class label distributions for producing more plausible pseudo bounding boxes, by a means to integrate the diffusion model into the teacher-student framework for SSL.

- We demonstrate state-of-the-art performance against existing methods on the ScanNet and SUN RGB-D benchmarks.

Thanks for your attention.