Learning Object-level Point Augmentor for Semi-supervised 3D Object Detection

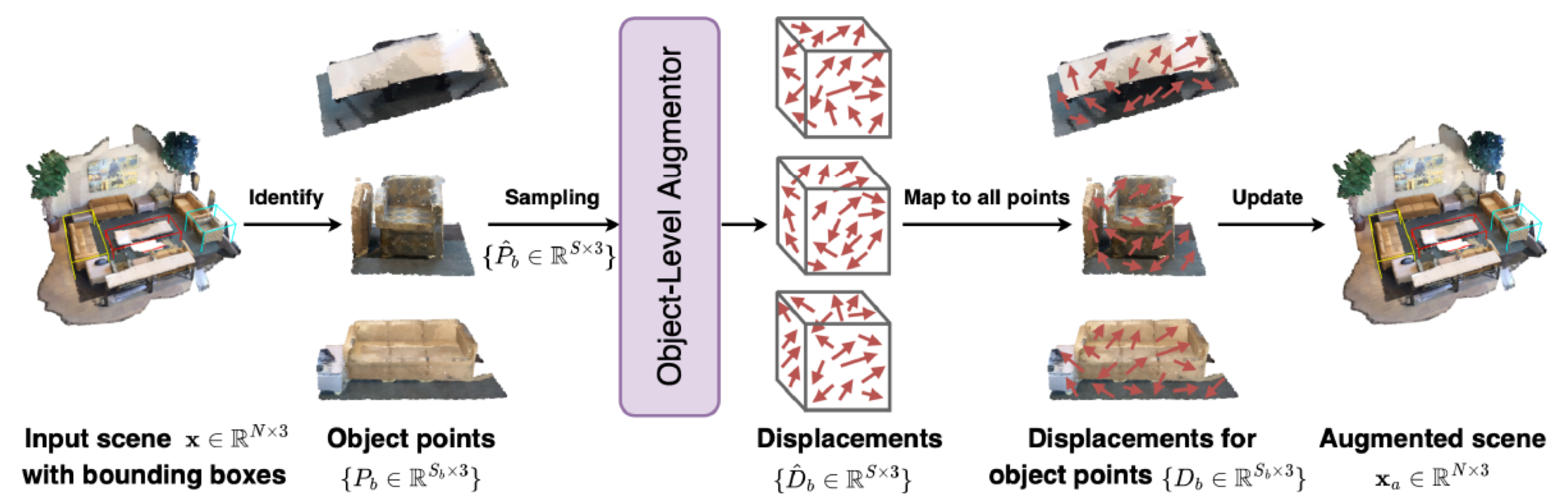
Cheng-Ju Ho[1] Chen-Hsuan Tai[1] Yi-Hsuan Tsai[2] Yen-Yu Lin[1] Ming-Hsuan Yang[3]
[1]National Yang Ming Chiao Tung University [2]Phiar Technologies [3]University of California at Merced

Observation

- Existing 3D semi-supervised methods usually employ only global augmentation, but it's sub-optimal
  - It ignores the object-level data variance, which is crucial for the instance-level object detection task.
- Apply augmentations to the point clouds within each object bounding box directly.
  - Its performance depends on proper augmentation settings.
- Compared with rotation, point displacement can enhance data variance while keeping object orientations.

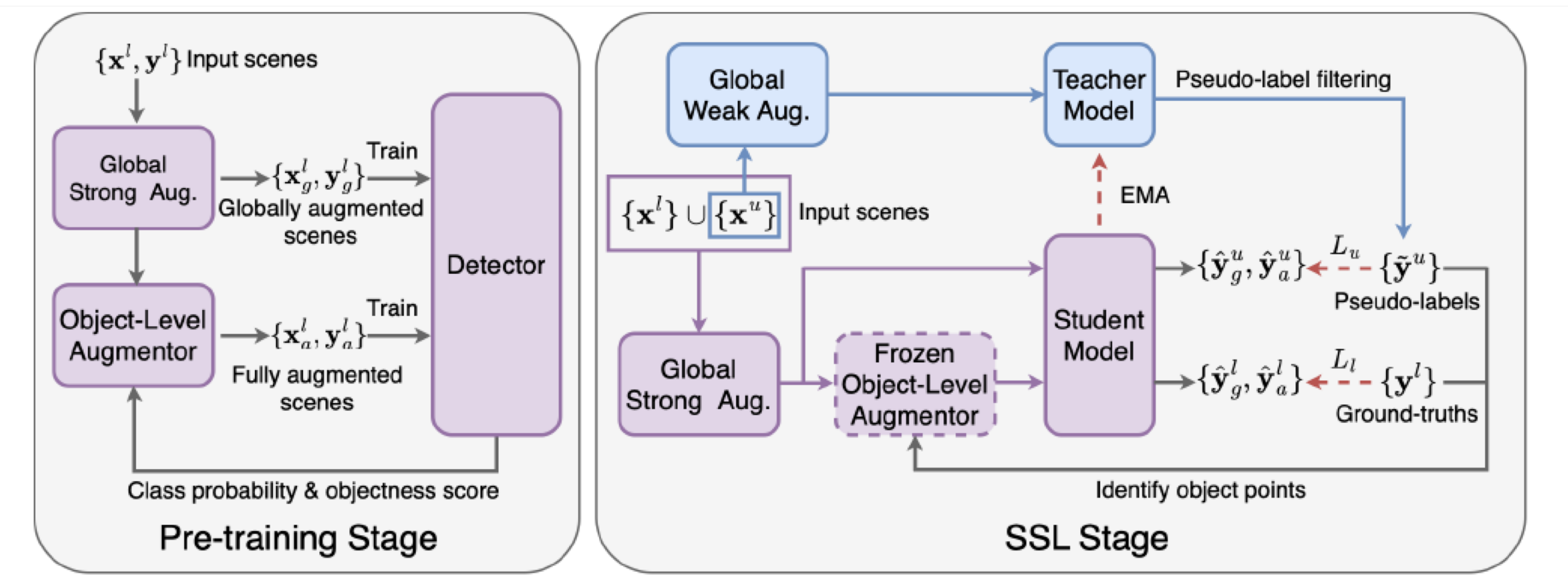
Setting	ScanNet 10%		SUN RGB-D 5%	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
Without Object-level Aug.	47.1	28.3	39.0	21.1
Pre-defined Object-level Aug. (scale, flip, rotation)	42.7	24.2	24.9	13.6
Pre-defined Object-level Aug. (displacement, range at 0.5%)	48.4	29.1	40.6	20.4
Pre-defined Object-level Aug. (displacement, range at 1%)	49.0	29.3	40.5	20.9
Pre-defined Object-level Aug. (displacement, range at 5%)	47.3	27.4	39.5	20.5

Object-level point augmentor



- Emphasize object instances rather than irrelevant backgrounds.
- Dynamically adjust the augmentation magnitude according to the detector's ability.
- After augmenting, making the augmented data more useful for object detector training.

Methodology



- Adversarial learning strategy.
  - With jointly pre-train a detector with an augmentor. The augmentor is optimized to generate proper augmented scene $\mathbf{x}_a$, while the detector is derived to localize and recognize the augmented data accurately.
- Augmentation Objective.
  - Augmented scene $\mathbf{x}_a$ should be more challenging. $\mathcal{L}_d(\mathbf{x}_a^l, \mathbf{y}_a^l) \geq \mathcal{L}_d(\mathbf{x}_g^l, \mathbf{y}_g^l)$
  - $\mathbf{x}_a$ and $\mathbf{x}_g$ should be classified as the same class.
  - Dynamical variable ρ makes $\rho\mathcal{L}_d(\mathbf{x}_g^l, \mathbf{y}_g^l)$ be the upper bound $\mathcal{L}_d(\mathbf{x}_a^l, \mathbf{y}_a^l)$

$$\mathcal{L}_A = \mathcal{L}_d(\mathbf{x}_a^l, \mathbf{y}_a^l) + \lambda \left|1 - \exp\left(\mathcal{L}_d(\mathbf{x}_a^l, \mathbf{y}_a^l) - \rho\mathcal{L}_d(\mathbf{x}_g^l, \mathbf{y}_g^l)\right)\right|$$

  - ρ is aware of objectness score and class probability.

$$\rho = \max(1, \exp(\hat{\mathbf{y}}_o \cdot \sum_{c=1}^{C} \hat{\mathbf{y}}_c \cdot \mathbf{y}_c))$$

- Teacher-Student Framework in SSL.
  - We initialize the student and teacher models from the pre-trained detector, and apply our object-level augmentor and asymmetric data augmentations to make this framework effective.

Quantitative results on indoor datasets

Dataset	Model	5%		10%		20%	
		mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5
ScanNet	VoteNet [16]	27.9±0.5	10.8±0.6	36.9±1.6	18.2±1.0	46.9±1.9	27.5±1.2
	SESS [36]	NA	NA	39.7±0.9	18.6	47.9±0.4	26.9
	3DIoUMatch [29]	40.0±0.9	22.5±0.5	47.2±0.4	28.3±1.5	52.8±1.2	35.2±1.1
	OPA	**41.9**±1.5	**25.0**±0.4	**50.5**±0.2	**32.7**±1.0	**54.7**±0.3	**36.8**±0.8
	Gain (%)	1.9↑	2.5↑	3.3↑	4.4↑	1.9↑	1.6↑
SUN RGB-D	VoteNet [16]	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 [36]	NA	NA	42.9±1.0	14.4	47.9±0.5	20.6
	3DIoUMatch [29]	39.0±1.9	21.1±1.7	45.5±1.5	28.8±0.7	49.7±0.4	30.9±0.2
	OPA	**41.6**±0.1	**23.1**±0.5	**47.2**±0.7	**29.6**±0.8	**50.8**±1.0	**31.5**±0.6
	Gain (%)	2.6↑	2.0↑	1.7↑	0.8↑	1.1↑	0.6↑

Source code:

Qualitative results on the ScanNet

Ground Truth OPA (ours) 3DIoUMatch